《计算方法丛书》编委会

序　　言

自然界和工程技术中的很多现象，例如自动控制系统的运行、电力系统的运行、飞行器的运动、化学反应的过程、生态平衡的某些问题等，其数学模型是常微分方程(组)的初值问题.很多偏微分方程也可以化为常微分方程组的初值问题求近似解.一般说来，常微分方程(组)的数值解法是比较成熟的，理论比较完整，也有很多方法可供选择．但是，有一类常微分方程组，在求数值解时遇到相当大的困难，这类常微分方程组解的分量有的变化很快，有的变化缓慢.常常出现这种现象：变化快的分量很快地趋于它的稳定值，而变化慢的分量缓慢地趋于它的稳定值．从数值解的观点看来，当解变化快时应该用小步长积分，当变化快的分量已趋于稳定，或者说已没有变化快的分量出现时，就应该用较大的步长积分.但是理论和实践都说明，很多方法，特别是显式方法的步长仍不能放大，否则便出现数值不稳定现象，即误差急剧增加，以致掩盖了真解，使求解过程无法继续进行.

常微分方程组的这种性质叫做刚性(Stiff)，这一问题近二十年来引起了很多计算数学工作者的重视，他们从理论上探讨这类问题的实质，并从各个角度寻求适用的数值解法．关于刚性已有了公认的数学定义，也建立了数值稳定性和稳定区域的概念，并且发表了大量数值解法的论文，取得了许多研究成果，在这里特别应当提到 G. Dahlquist, J. C. Butcher, C. W. Gear 等人，他们都做了重要的工作.

六十年代初，在我们的工作中也发现了这个问题．我们和其他单位的同志都做过一些工作．本书第一、十一、十二这三章主要是我们多年来在实际工作中的一些经验的总结.

刚性常微分方程组数值解法的研究还有很多问题尚待解决.

非线性问题数值稳定性的研究刚开始，在解各种实际问题，特别是复杂的大系统问题时，需要有更适用的算法，同时也需要研究各种相关的问题.

这本书介绍了刚性常微分方程组初值问题数值解法的一些基本的研究成果和实际构造算法的思路，并给出一些研究和构造算法的背景材料，以便于实际解题的技术人员参考，本书也可作为在这个领域开始工作的计算数学和数学工作者的读物.

本书共十三章. 第一章是对刚性常微分方程的概念和常用的数值稳定性定义的总的叙述，第二章介绍解刚性常微分方程线性多步法的稳定性理论. 这两章是阅读本书的基础. 第四章介绍了与刚性常微分方程数值解法研究有关的 Padé 近似的一些处理方法和结果. 第三章、第五章至十三章均是介绍刚性常微分方程数值解法和处理的思想.

本书利用了国内外许多作者的材料，审校者和出版社的同志也提出了许多很好的意见，在此表示感谢. 由于作者水平所限，缺点和错误在所难免，欢迎读者批评指正.

作者

1984 年春

目　　录

第一章 引 论

本章介绍刚性常微分方程的概念，描述用数值方法解这种方程时遇到的困难．我们还将列出对刚性方程的数值解法进行分析时常用的一些稳定性定义，并给出几个不同学科中出现的刚性方程的简单的例子．最后叙述稳定性区域的计算方法．

§1 刚性常微分方程

在可以用常微分方程来描述的许多实际的物理或化学过程中，往往包含许多复杂的子过程及它们之间的相互作用，其中有的子过程表现为快变化的，而另一些相对来说是慢变化的，并且变化速度可以相差非常大的量级．相应地，描述这些过程的常微分方程的解中也将包含快变分量和慢变分量．如果在一个过程中的快变子过程与慢变子过程的变化速度相差非常大，在数学上称这种过程具有刚性 (Stiff)，而描述这种过程的常微分方程称为刚性方程．

为精确地描述过程的刚性性质，考虑线性系统

$$\frac{dy(t)}{dt} = Ay(t) + \phi(t), \tag{1.1}$$

其中 $y(t) = (y_1(t), y_2(t), \cdots, y_m(t))^T$ 是待求的 t 的 m 维向量函数，而 $\phi(t) = (\phi_1(t), \phi_2(t), \cdots, \phi_m(t))^T$ 是已知的向量函数，t 是独立变量，可看成是时间，A 是 $m \times m$ 矩阵．不失一般性，假定矩阵 A 的 Jordan 标准型是对角矩阵，其特征值为

$$\lambda_k = \alpha_k + i\beta_k,\ k = 1, 2, \cdots, m,$$

相应的特征向量记成 ξ_k．(1.1) 的解有形式

$$y(t)=\sum_{k=1}^{m} c_k e^{\lambda_k t}\xi_k+\phi(t). \tag{1.2}$$

假定 $\mathrm{Re}\lambda_k<0$, $k=1, 2,\cdots, m$, 即(1.1)是渐近稳定的. 当 $t\to+\infty$ 时,有 $\sum_{k=1}^{m} c_k e^{\lambda_k t}\xi_k\to 0$, 则称 $\sum_{k=1}^{m} c_k e^{\lambda_k t}\xi_k$ 为(1.1)的暂态解,而称项 $\phi(t)$ 为稳态解. 各个 $e^{\lambda_k t}$ 称作(1.1)的齐次方程组的解分量(简称解分量).

暂态解 $\sum_{k=1}^{m} c_k e^{\lambda_k t}\xi_k$ 可以表示成解分量

$$u_k(t)=e^{\lambda_k t}=e^{\alpha_k t}e^{i\beta_k t} \quad k=1,\cdots,m$$

的线性组合. 实部 α_k 确定 $u_k(t)$ 的衰减特性, 而虚部 β_k 确定这个量的振荡特性. 对于一个稳定系统, λ_k 的实部 α_k 一定是负的. 工程上称量 $\tau_k=-1/\alpha_k$ 为时间常数,用它来表征量 $u_k(t)$的衰减速度. 因为每经过时刻 τ_k, $e^{\alpha_k t}$ 衰减 e^{-1} 倍, 即约 $\frac{1}{3}$ 倍, $-\alpha_k$ 越大衰减越快. $e^{i\beta_k t}$ 的振荡频率为 $\beta_k/2\pi$, β_k 越大,振荡越快. 一般 $u_k(t)$ 是衰减的或振荡衰减的,但各个 $u_k(t)$ 的衰减速度之间的差异可能是很大的,这由刚性比来刻划.

定义 1.1 (Lambert[67]) 线性系统(1.1)称作是刚性方程,如果有

(i) $\mathrm{Re}\lambda_k<0$, $k=1, 2,\cdots, m$,

(ii) $r=\max\limits_{k=1,\cdots,m}|\mathrm{Re}\lambda_k|/\min\limits_{k=1,\cdots,m}|\mathrm{Re}\lambda_k|\gg 1$,

比值 r 称作刚性比.

根据这个定义,我们来描述刚性方程所具有的一些性质.

(i) 刚性方程是渐近稳定的, 解曲线从不同的初值都将趋向于它的稳态解. 但各个解分量的衰减特性是不同的. 衰减快的称作快变分量,衰减慢的称作慢变分量. 刚性方程的解曲线将很快衰减到由慢变分量所确定的解曲线上. 例如,考虑二阶齐次方程

$$\mu\frac{d^2y}{dt^2}+\frac{dy}{dt}+y=0.$$

它的解有形式 $y(t)=c_1e^{\lambda_1 t}+c_2e^{\lambda_2 t}$ 其中 c_1 和 c_2 是二个任意常数,由初始条件确定,λ_1 和 λ_2 是它的特征方程的根,即

$$\lambda_1 \approx -\frac{1}{\mu}+1+\mu,$$

$$\lambda_2 \approx -1-\mu.$$

当 $\mu>0$ 很小时,$y(t)$ 将很快地衰减到解曲线

$$\tilde{y}(t)=c_2e^{\lambda_2 t}.$$

因此刚性方程的解曲线可分为二段,开始的一段为快变段,解曲线中的快变部分迅速地衰减到可忽略的程度,将快变段称作边界层,其所经历的时间记作 τ_{BL},一般它取为快变分量衰减到原来的 $\frac{1}{20}$ 时的时间,即取 $\tau_{BL}\approx 3/|\mathrm{Re}\lambda_i|$. 另外一段是慢变段,或称边界层外的段,它由 $|\mathrm{Re}\lambda_i|$ 较小的解分量来刻划. 当然上面将解曲线的分段是相对的,例如,对于特征值为

$$\lambda_1=-10^5,\ \lambda_2=-10^3,\ \lambda_3=-1$$

的三个方程的线性方程组,对 λ_2 来说 $\tau_{BL_2}=3\cdot10^{-3}$,而对 λ_1 来说 $\tau_{BL_1}=3\cdot10^{-5}$.

(ii) 刚性方程组具有奇异摄动的性质. 由于解曲线中的快变部分在边界层内很快地衰减掉,在边界层外,(1.1)的解曲线中所含的量的个数减少,使得解曲线的各个坐标之间不再是线性独立的,而存在若干个代数关系式. 利用这些关系式,可以用低阶的方程组代替(1.1). 为了说明这个性质,考虑刚性方程组

$$\begin{aligned}\frac{dy^{(1)}}{dt}&=-501y^{(1)}+500y^{(2)}\\ \frac{dy^{(2)}}{dt}&=500y^{(1)}-501y^{(2)}.\end{aligned}\tag{1.3}$$

它的解为

$$y^{(1)}(t)=0.5[y^{(1)}(0)-y^{(2)}(0)]e^{-1001t}+0.5[y^{(1)}(0)+y^{(2)}(0)]e^{-t},$$

$$y^{(2)}(t)=-0.5[y^{(1)}(0)-y^{(2)}(0)]e^{-1001t}+0.5[y^{(1)}(0)$$

$$+y^{(2)}(0)]e^{-t}.$$

在边界层外近似有等式

$$y^{(1)}(t)=y^{(2)}(t),\tag{1.4}$$

利用这个等式，可以将二阶方程(1.3)降阶为微分方程

$$\frac{dy^{(2)}}{dt}=-y^{(2)}$$

和代数方程 (1.4) 的代数微分方程组．这种性质在导数前含有小参数的微分方程组

$$\begin{aligned}\mu\frac{dx}{dt}&=f(t,x,y),\\ \frac{dy}{dt}&=g(t,x,y)\end{aligned}\tag{1.5}$$

中是常见的．(1.5) 是刚性方程最早的例子之一．当 $\mu\to 0$ 时，(1.5)的解将收敛到退化组

$$\begin{aligned}f(t,\bar{x},\bar{y})&=0,\\ \frac{d\bar{y}}{dt}&=g(t,\bar{x},\bar{y})\end{aligned}\tag{1.6}$$

的解，可以利用刚性方程组的奇异摄动性质来构造计算方法．

(iii) 定义1.1实际上表明矩阵 A 是病态的．

对于非线性系统

$$\frac{dy}{dt}=f(t,y)\tag{1.7}$$

的刚性性质可以按下面的方式定义． 令 $\tilde{y}(t)(t\in[a,b])$为(1.7)的满足初始条件 $\tilde{y}(a)=y_0$ 的精确解．我们在解 $\tilde{y}(t)$ 的邻域中来考察(1.7)的解的特性．在这个邻域中，(1.7)可以用线性摄动方程

$$\frac{dy}{dt}=J(t)(y-\tilde{y}(t))+f(t,\tilde{y}(t))$$

或者

$$\frac{dy}{dt}=J(t)y+[f(t,\tilde{y}(t))-\tilde{y}(t)]\tag{1.8}$$

来近似，其中 $J(t)$ 是在点 $(t, \tilde{y}(t))$ 处计算 $f(t, y)$ 的 Jacobi 矩阵 $\partial f(t, y)/\partial y$ 的值. 如果 $J(t)$ 在区间 $[a, b]$ 上的变化充分小，则在这个区间上它可以用某一个固定的 $J(t_0)$ 代替，$t_0 \in [a, b]$. 作这样的替代后，(1.8)具有(1.1)的形状. 因此，在 $[a, b]$ 上(1.7)的解可以近似地表成

$$y(t) \approx \tilde{y}(t) + \sum_{i=1}^{m} c_i e^{\lambda_i t} \xi_i, \tag{1.9}$$

其中 c_i 是常数，λ_i 和 ξ_i 分别是 $J(t_0)$ 的特征值和相应的特征向量，$i = 1, \cdots, m$. 这样，只要在定义 1.1 中用 $J(t_0)$ 的特征值代替 (1.1) 中的矩阵 A 的特征值，我们就可以定义 (1.7) 在区间 $[a, b]$ 上的刚性性质. 这表示(1.7)的刚性性质可以由它的右函数的 Jacobi 矩阵的特征值来定义，这时定义 1.1 中的刚性比是依赖于 t_0 的，可以称作局部刚性比. 但是，非线性微分方程的解的相态是非常复杂的，这样定义的刚性比并不象对线性方程那样一定能刻划解曲线所具有的性质. 为了说明，只需考虑三阶变系数线性方程组

$$\frac{dy}{dt} = A(t)y, \tag{1.10}$$

其中当 $t \in [0, 1]$ 时

$$A(t) = \begin{bmatrix} -1 + 100\cos 200t & +100(1 - \sin 200t) & 0 \\ -100(1 + \sin 200t) & -(1 + 100\cos 200t) & 0 \\ 1200(\cos 100t + \sin 100t) & 1200(\cos 100t - \sin 100t) & -501 \end{bmatrix},$$

这时矩阵 $A(t)$ 的特征值为常值

$$\lambda_1 = -501, \quad \lambda_2 = -1, \quad \lambda_3 = -1.$$

按照上述刚性微分方程的定义，(1.10)在$[0, 1]$上是刚性方程组，但实际上(1.10)是不稳定的，其解有形式

$$y^{(1)}(t) = c_1 e^{99t} \cos 100t + c_2 e^{-101t} \sin 100t,$$
$$y^{(2)}(t) = -c_1 e^{99t} \sin 100t + c_2 e^{-101t} \cos 100t,$$
$$y^{(3)}(t) = 2c_1 e^{99t} + 3c_2 e^{-101t} + c_3 e^{-501t},$$

常数 c_1, c_2, c_3 由初始条件确定．对于大多数工程问题，用局部 Jacobi 矩阵的特征值可以刻划非线性常微分方程的稳定性，这方面的结果可参见秦元勋的[10]或许淞庆的[13]．

在实际问题中，刚性比可高达 10^6 以上的量级．如果 r 的量级为 10，称作是临界刚性的．

刚性方程在文献中也称作病态方程或坏条件方程，具有差别大的时间常数问题或具有大的 Lipschitz 常数的问题．这种类型的问题在控制系统工程，电子网络，生物学，物理及化学动力学过程中经常遇到．例如，在控制系统中，控制部件一般反应灵敏，是快变的，具有小的时间常数，而受控物体一般惯性大，是慢变的，具有大的时间常数．对于宇宙航行中的运载器，通常是通过控制部件来控制质心的运动．姿态运动是快变的，而质心运动是慢变的．在多成分的化学反应中，有些反应速度快，几乎瞬时就达到稳定状态；而有些反应速度慢，两者的差别可以有好几个量级．对于复杂的电子网络，由于不同的寄生电容的影响，时间常数也有很大的差异．在传热、扩散、分馏等过程中，把分布参数离散化成集中参数而得到的常微分方程组经常是刚性方程．在 §3 中，我们从一些不同的学科选取几个例子来加以说明．

由定义 1.1 可以看到，刚性性质是数学问题本身的性质．它不依赖于求解这个问题的数值方法．但是正是由于这个性质，使得传统的常微分方程的数值积分方法遇到极大的困难．为了克服这个困难，刚性常微分方程数值积分方法的研究成为数值方法中最活跃的方向之一．

为了说明传统的数值方法求解刚性方程所遇到的困难，首先把问题做一些简化．假定(1.1)的矩阵 A 的 Jordan 标准型是对角矩阵 $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$，即有 $m\times m$ 矩阵 S，使

$$S^{-1}AS = \Lambda.$$

将 S^{-1} 左乘(1.1)，并令 $z = s^{-1}y$，得

$$\frac{dz}{dt} = \Lambda z + S^{-1}\phi(t). \tag{1.11}$$

这表示若考虑 y 的某个线性变换的象 $z=S^{-1}y$ 时，z 的 m 个分量各满足一个独立的标量方程

$$\frac{dz^{(k)}}{dt}=\lambda_k z^{(k)}+\phi_k(t),\ (k=1,2,\cdots,m), \quad (1.12)$$

$\phi_k(t)$ 为 $S^{-1}\phi(t)$ 的第 k 个分量. 而 $y=Sz$.

假设 y 的初值 y_0 有一个误差 Δy_0. 显然(1.1)的解 y 的误差 Δy 满足方程

$$\frac{d\Delta y}{dt}=A\Delta y,$$

即 Δy 满足(1.1)的齐次组

$$\frac{dy}{dt}=Ay, \quad (1.13)$$

只是初值取 Δy_0. 同理可得误差 Δy 的象 $\Delta z=S^{-1}\Delta y$满足

$$\frac{dz}{dt}=\Lambda z. \quad (1.14)$$

用 Euler 法解(1.1)得到的差分方程组是

$$y_{n+1}=y_n+h(Ay_n+\phi(t_n)). \quad (1.15)$$

假设在某一步有一个误差,以后计算不再有误差,考虑 y_n 的误差 Δy_n 变化的情况. 不失一般性,假定这个误差是初值 y_0 的误差,用同一个符号 Δy_0 表示它. 显然 Δy_n 满足(1.15)的齐次组

$$y_{n+1}=y_n+hAy_n, \quad (1.16)$$

其初值是 $y_0=\Delta y_0$, 即是用 Euler 法解齐次组(1.13)所得的差分方程组.

类似地取变换 $z_n=S^{-1}y_n$, (1.16)变成

$$z_{n+1}=z_n+h\Lambda z_n, \quad (1.17)$$

初值 $z_0=S^{-1}\Delta y_0$, 这相当于用 Euler 法解(1.14)所得的差分方程组. (1.14)的 m 个方程

$$\frac{dz^{(k)}}{dt}=\lambda_k z^{(k)},\ k=1,\cdots,m,$$

是相互独立的,(1.17)中的 m 个方程

$$z_{n+1}^{(k)} = z_n^k + h\lambda_k z_n^{(k)},\ k = 1, 2, \cdots, m,$$

也互相独立,将上述的处理方法应用到其它的数值积分方法上,也可得到类似的结果. 因此,研究(1.1)的数值积分中误差的变化,在很大程度上可以归结为用同一个方法研究标量方程初值问题

$$\frac{dy}{dt} = \lambda y,\ y(0) = y_0 \tag{1.18}$$

的解的性质. 因为工程问题要求 $\text{Re}\lambda < 0$, 所以我们也做这样的假定. 这个微分方程叫做试验方程,许多计算稳定性的定义都以它为基础. 一般说,当 n 无限增大时,误差无限增加的数值方法是不可用的,而误差趋于零的数值方法是可用的. 由于上面的对非线性系统的线性摄动理论的考虑,试验方程对于误差变化即计算稳定性的研究具有很广泛的代表性.

考虑用固定步长的数值方法求解问题(1.18). 令 $h > 0$ 为数值积分的步长, $t_i = ih$, $i = 0,1,\cdots$ 为积分的节点, y_i 为解得的精确解 $y(t_i)$ 的近似值.

定义 1.2 用一个数值积分方法以定步长 h 解试验方程(1.18),当 $n \to \infty$ 时, $y_n \to 0$, 则称用步长 h 的这个数值积分过程为计算稳定的,或简称稳定的,否则称为计算不稳定的.

应当注意,这里定义的是数值积分的计算过程的稳定性. 当计算步骤增加时,误差的趋势说明了数值积分方法的可用性. 这与物理上描写的运动稳定性是不同的概念.

设应用的数值积分方法是 Euler 公式,则在近似解 y_i $(i = 0,1,\cdots)$之间有递推式

$$y_{i+1} = (1 + h\lambda)y_i. \tag{1.19}$$

计算开始时,为了使 y_i 能精确地近似 $y(t_i)$, 从精确度的角度要求步长 h 适当的小. 但是当 $|y_i|$ 已充分小,可以近似地认为已达到(1.18)的稳定状态时,续续积分(1.18)就不必要求 h 很小. 这时积分精度不重要了,只需从数值积分的稳定性要求步长 h 满足不等式

$$|1 + h\lambda| < 1. \tag{1.20}$$

满足不等式(1.20)的量 $q=h\lambda$ 是复平面上的一个有限区域中的点，即属于以−1为中心，以1为半径的圆的内部，见图1.1.这是对步长的一个约束，在整个数值积分过程中，选取的步长 h 应使 $h\lambda$ 不越出图1.1中给出的有限区域.

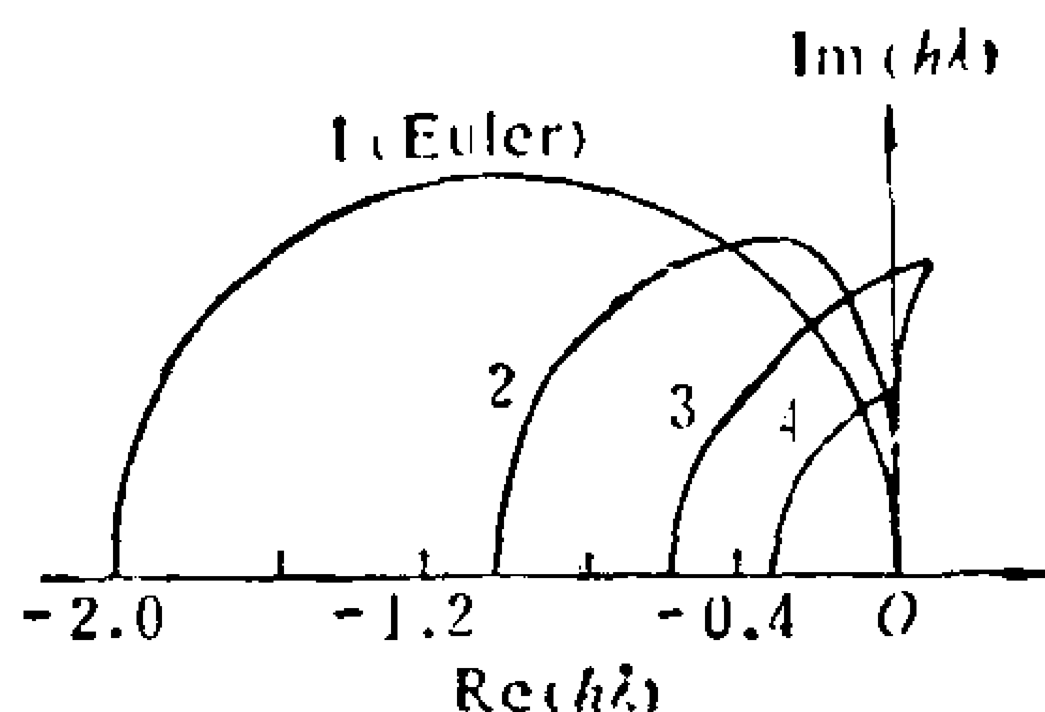


图 1.1 Euler 公式的稳定区域（只画上平面）

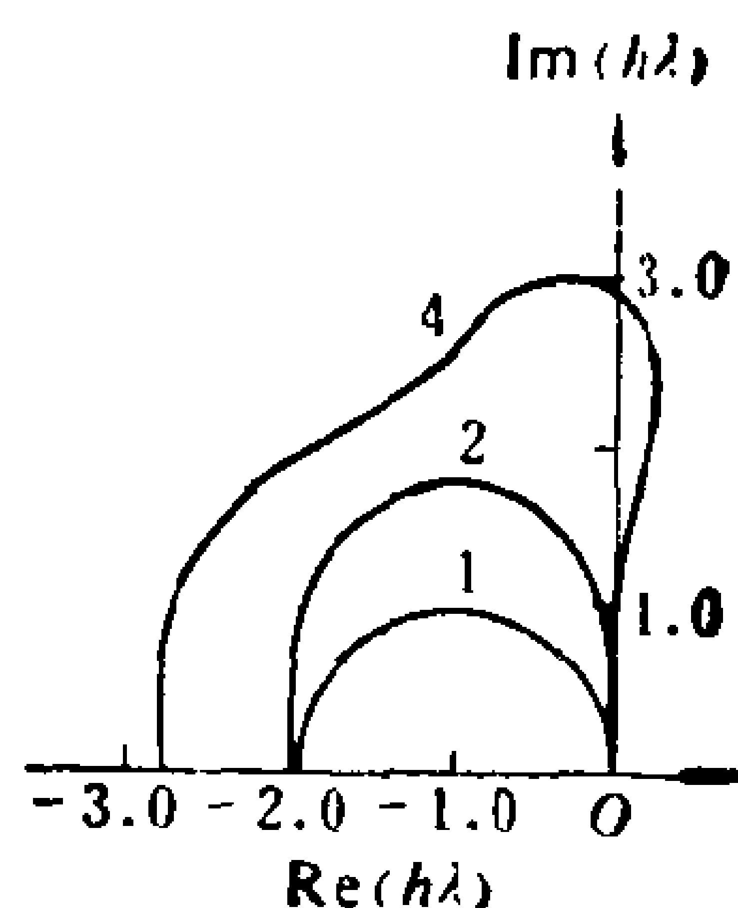


图 1.2 Runge-Kutta 公式的稳定区域（只画上平面）

类似地，若应用的数值积分方法是四阶显式 Runge-Kutta 方法，得到递推式

$$y_{i+1}=\left[1+h\lambda+\frac{1}{2}(h\lambda)^2+\frac{1}{6}(h\lambda)^3+\frac{1}{24}(h\lambda)^4\right]y_i, \tag{1.21}$$

选取的步长 h 必须满足不等式

$$\left|1+h\lambda+\frac{1}{2}(h\lambda)^2+\frac{1}{6}(h\lambda)^3+\frac{1}{24}(h\lambda)^4\right|<1, \tag{1.22}$$

满足不等式(1.22)的量 $q=h\lambda$ 也组成复平面上的一个有限区域，即图1.2闭曲线的内部.

将上述的有限区域分别称作 Euler 法和 Runge-Kutta 方法的稳定区域. 对于一般的递推的数值积分公式，我们引进下面的定义.

定义 1.3 数值积分公式的稳定区域 R 为集合

$R=\{h\lambda|$ 以固定的步长 $h>0$ 将公式应用到方程

$$\frac{dy}{dt} = \lambda y,\ y(t_0) = y_0,\ \mathrm{Re}\lambda < 0,$$

所得到的序列 $\{y_i\}$，当 $i \to \infty$时有 $y_i \to 0\}$.

除 Euler 法和四阶显式 Runge-Kutta 公式外，一些其它的常用数值积分公式的稳定区域也是复平面上的有限区域，例如 Adams-Bashforth 公式和 Adams-Moulton 公式均是这样的。特别若 λ 是负实数时，[19]列出了步长 h 必须满足的不等式，

Euler 公式	$\|h\lambda\| < 2$,
二点 Adams-Bashforth 公式	$\|h\lambda\| < 1$,
三点 Adams-Bashforth 公式	$\|h\lambda\| < 0.55$,
四点 Adams-Bashforth 公式	$\|h\lambda\| < 0.3$,
三点 Adams-Moulton 公式	$\|h\lambda\| < 2.4$,
四点 Adams-Moulton 公式	$\|h\lambda\| < 24/9$,
四阶显式 Runge-Kutta 公式	$\|h\lambda\| < 2.785$.

由此可以看到，大多数常用的传统数值积分方法所用的步长与 $|\lambda|$ 有着密切的关系. $|\lambda|$ 愈大，选取的积分步长应愈小. 正是由于这种现象，使得用传统的数值积分方法来求解刚性方程遇到极大的困难.

当求刚性方程(1.1)的数值解时，近似解中将包含各个 $e^{\lambda_k t}$ 的近似. 这相当于在积分(1.1)的同时，求解 m 个标量方程(1.18)，其中 λ 分别取 λ_k. 当数值积分时，对应于$-\mathrm{Re}\lambda_k$ 大的方程，其近似解暂态成分 $e^{\lambda_k t}$ 将很快衰减到可以忽略的程度，即认为已达到对应方程的稳态解. 往后的积分可以看成是求对应于$-\mathrm{Re}\lambda_k$ 为小的(1.18)的稳态解. 此时，(1.1)的数值解的量值将由稳态解的近似和对应于$-\mathrm{Re}\lambda_k$ 为小的方程(1.18)的近似解来确定. 若要积分到稳态解，则积分区间的长度的量级为 $1/\min\limits_{i=1,\cdots,m}(-\mathrm{Re}\lambda_i)$. 若还需要计算稳态解的性态，则积分区间长度的量级还要大一点. 但是不管每个暂态分量 $e^{\lambda_k t}$ 的近似值对(1.1)的近似解是否起作用，对(1.1)进行数值积分的同时必须看成对每个 $\lambda = \lambda_k$ 的(1.18)也

进行数值求解．这时选取的步长 $h>0$ 仍应使所有的 $h\lambda_k(k=1,\cdots,m)$ 均在方法所对应的稳定区域中．于是积分步长将由

$$\max_{i=1,\cdots,m}(-\mathrm{Re}\lambda_i)$$

来确定，步长的量级为 $1/\max_{i=1,\cdots,m}(-\mathrm{Re}\lambda_i)$．由上面的分析，对于具有有限稳定区域的数值积分方法，用固定步长求刚性方程 (1.1) 的数值解时，总的积分步数的量级将不小于

$$\max_{i=1,\cdots,m}(-\mathrm{Re}\lambda_i)/\min_{i=1,\cdots,m}(-\mathrm{Re}\lambda_i),$$

它恰好是(1.1)的刚性比．当刚性比很大时，用传统的数值积分方法解刚性方程的花费就太大了，有时甚至是无法实现的．事实上，上述步长的选取是不合理的，即上述的步长选取是由对解不起作用的成分确定的．正是为了克服这个不合理性，才开创了刚性方程数值方法的研究领域．

定义 1.1 是对方程组定义刚性性质，它不包含单个方程的情形，也不包含方程组中具有实部为零或实部为很小的正数的情形．按照 Shampine 和 Gear(1979) 的观点，若线性系统(1.1)满足下述三个条件，则称作刚性方程，(i) 矩阵 A 的所有特征值的实部不是很大的正数的值，(ii) A 至少有一个特征值的实部是很大的负数，(iii)对应于具有最大负实部的特征值的解分量变化是缓慢的．这种定义将包含上述的定义 1.1 所缺少的情形．按这种观点，当对应于具有最大负实部的特征值的解分量变化很快时（即处在边界层内），线性系统(1.1)不称作刚性方程，而当这解分量衰减成很小的量时才称作刚性方程．这表示一个问题在自变量的一些区间中可以是刚性问题，而在另外一些区间中可以不是刚性问题． 这种提法反映了实际计算的状况．因在边界层内，为精确地确定解，需要较小的步长，而在边界层外，解分量已变得充分小，这时刚性性质才变成计算中需要克服的困难．

本书中的刚性方程的概念将兼顾这两种定义．即定义 1.1 和上述的 Shampine 和 Gear 的观点．

§2 常用的稳定性定义

由 §1 看出，用传统的数值积分方法求解刚性方程遇到困难的主要原因是虽然对应于大的$-\mathrm{Re}\lambda_i$ 的量，$e^{\lambda_i t}$ 在近似解中已不起作用，但由于稳定区域是有限的，为保持相应方程（1.18）的数值积分的稳定性，步长仍将由$-\mathrm{Re}\lambda_i$大的成分来确定．鉴于这一点，希望研究具有无限稳定区域的方法．本节给出具有无限稳定区域的一些常用的稳定性定义．

Dahlquist（1963）提出A稳定性的概念[48]．

定义 1.4 数值积分公式称作A稳定的，如果这个公式的稳定区域含有复开左半平面．

用数值积分公式来求刚性方程的数值解时，我们自然希望它是A稳定的，因为它表示公式对任意大的步长均是稳定的，稳定性的要求对步长 h 并未构成任何限制．当大的$-\mathrm{Re}\lambda_i$ 的量 $e^{\lambda_i t}$ 衰减到很小时，计算的步长可以由$-\mathrm{Re}\,\lambda_i$ 小的项来选取．因此每构造一个求解刚性方程的数值积分公式时，首先应对其进行A稳定性分析．但是，Dahlquist 在引入A稳定性的同时，证明了一个具有限制性的结果：显式的线性多步法（包括显式 Runge-Kutta 方法)不可能是A稳定的；A稳定的隐式线性多步法的阶不能超过 2，而在所有A稳定的二阶方法中，梯形公式具有最小的局部截断误差常数(这个经典性的结果我们在第二章中给出)．由于 Dahlquist 这一开创性结果的推动，并且为了突破这个结果的限制，刚性方程的数值方法的研究主要朝着两个方面发展．一个方面是减弱A稳定性的定义，也就是说要求方法的稳定区域不必包含整个开左半平面，而只需包含左半平面的一部分．这样，先后由 Widlund (1967) 引进 $A(\alpha)$ 稳定性，由 Gear（1968）引进刚性稳定性，由 Cryer (1973) 引进 A_0 稳定性．构造的具有这些稳定性性质的方法仍适合一大类刚性问题的数值求解，它们的最高精度阶可以超过 2．在这一节中，我们将叙述这些稳定性的定义．另外一个发

展的方面是保持 A 稳定性的定义，而构造数值积分方法的新的形式，来突破 Dahlquist 给出的限制．本书将在后面的章节中介绍一些这方面的进展．

在具体求解刚性方程(1.1)时，特征值 $\lambda_j(j=1,\cdots,m)$ 都是固定的，而在求解 Jacobi 矩阵 $\partial f(t,y)/\partial y$ 变化缓慢的非线性系统(1.7)时，相应的矩阵特征值 λ_i 的变化也是缓慢的，$h\lambda_i$ 的变化区域包含在图 1.3 的一组楔形中．因此，为了能有效地数值求解，只要稳定区域能包含这些楔形就行了．为此，Widlund(1967) 定义了 $A(\alpha)$ 稳定性[113]．

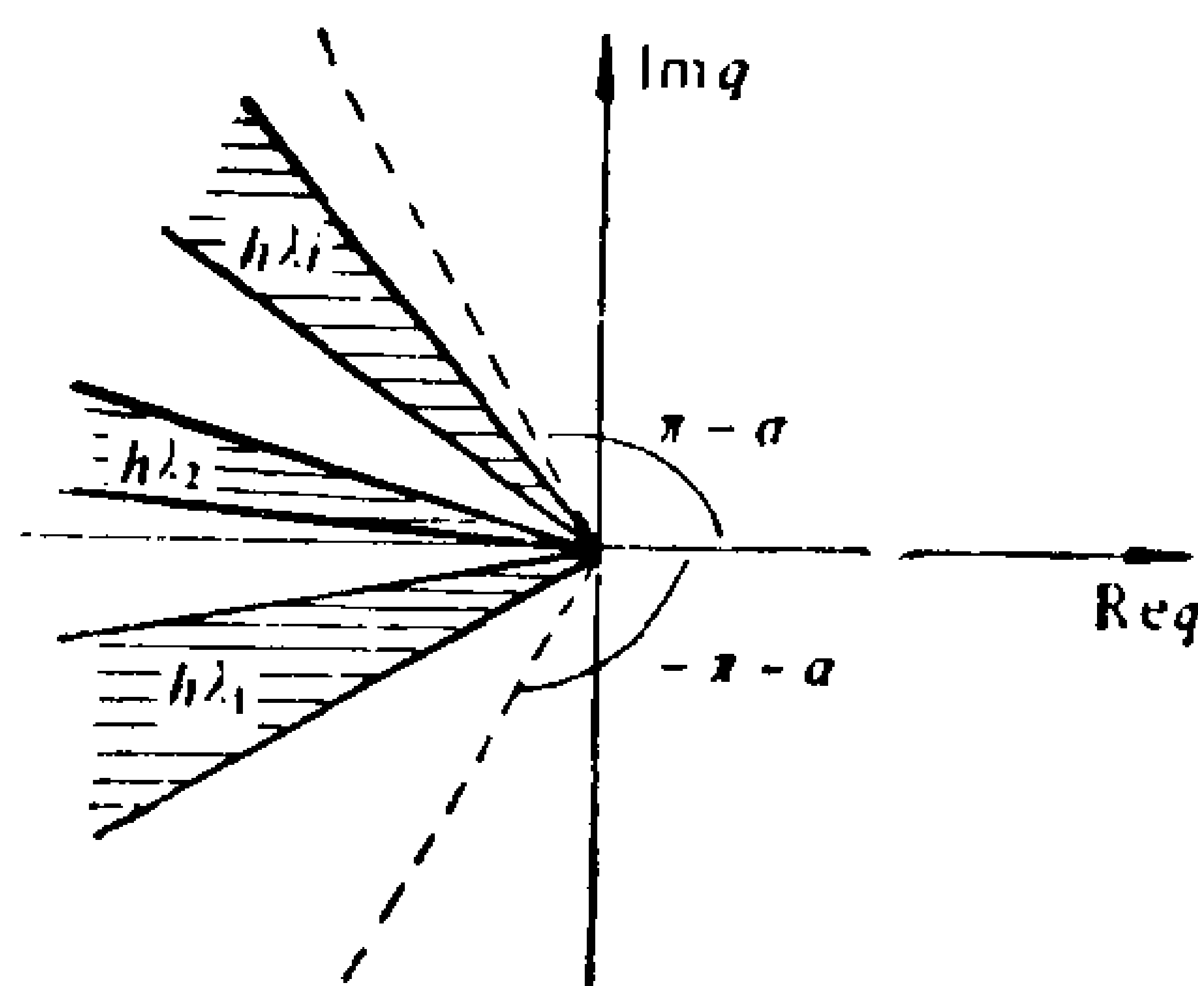


图 1.3 $A(\alpha)$ 稳定性区域

定义 1.5 数值积分公式称作是 $A(\alpha)$ 稳定的，$\alpha\in\left(0,\frac{\pi}{2}\right)$，如果方法的稳定区域包含 $h\lambda$ 复平面上的集合

$$\{h\lambda\,|\,|\arg(-\lambda)|<\alpha,\ |\lambda|\neq 0,\ h>0\}.$$

数值积分公式称作是 $A(0)$ 稳定的，如果存在充分小的 $\alpha\in\left(0,\frac{\pi}{2}\right)$，公式是 $A(\alpha)$ 稳定的．数值积分公式称作是 $A\left(\frac{\pi}{2}\right)$ 稳定的，如果它对所有 $\alpha\in\left(0,\frac{\pi}{2}\right)$ 是 $A(\alpha)$ 稳定的．

$A\left(\frac{\pi}{2}\right)$ 稳定性就是 A 稳定性．

Cryer (1973)[46] 引进了 A_0 稳定性．

定义 1.6 数值积分公式称作是 A_0 稳定的，如果公式的稳定区域包含整个负实轴.

将这定义与定义 1.4 和定义 1.5 相比较，立即可以看出，A 稳定或者 $A(\alpha)$ 稳定的公式一定是 A_0 稳定的. 即 A_0 稳定的公式类包含 A 稳定的公式类和 $A(\alpha)$ 稳定的公式类. 通过详细研究 A_0 稳定的公式的性质，将能够确定适合于求解刚性方程的数值方法必须具备的一些条件. 这是研究 A_0 稳定性的理论意义. 另一方面，许多实际的刚性问题可能只具有实的特征值，于是 A_0 稳定的数值积分公式对于求解一些特殊的刚性方程是有效的.

$A(0)$ 稳定性与 A_0 稳定性是不同的. 事实上 A_0 稳定的公式类将包含 $A(0)$ 稳定的公式类，并且二者是不重合的.

上面的定义只考虑稳定性. Gear(1968)定义的刚性稳定性既考虑了稳定性，又考虑了数值近似的精度. 对于正数 D, a 和 θ, 定义集合

$$R_1 = \{h\lambda \mid \mathrm{Re}(h\lambda) \leqslant -D\},$$

$$R_2 = \{h\lambda \mid -D < \mathrm{Re}(h\lambda) < a, |\mathrm{Im}(h\lambda)| < \theta\}.$$

定义 1.7 数值积分公式称作是刚性稳定的，如果它是收敛的，并且存在正常数 D, θ, a, 集合 R_1 含在公式的稳定区域内，而公式在区域 R_2 上具有高的精度，并具有相对或绝对的稳定性.

Gear 提出这种定义的想法是这样的. 微分方程(1.1)或(1.7)的解中含有形式为 $e^{\lambda t}$ 的成分，用数值方法以步长 h 积分一步时，这种量改变大约 $e^{\lambda h}$ 倍，如果 $\lambda h = u + iv$，则改变的幅度为 e^u，如果 $u < -D$，则从量值上至少减少到原来的 e^{-D} 倍. 当 D 适当大时，在区域 R_1 中这种量的绝对值将很快减小到可忽略的程度，因而在该区域中，公式的积分精度可少考虑，仅需要保证方法是稳定的. 在含原点的区域 R_2 中，为了求得 $e^{\lambda t}$ 的比较精确的近似值，数值积分公式的精度和稳定性均是需要考虑的. 对于复平面上的其它部分，在定义 1.7 中未对公式提出要求，这是因为如果 $u > a > 0$，每积分一步，量 $e^{\lambda t}$ 至少增加 e^a 倍，从而必须选取足够小的步长以便能够适应这种变化. 实际上，是要选取 h

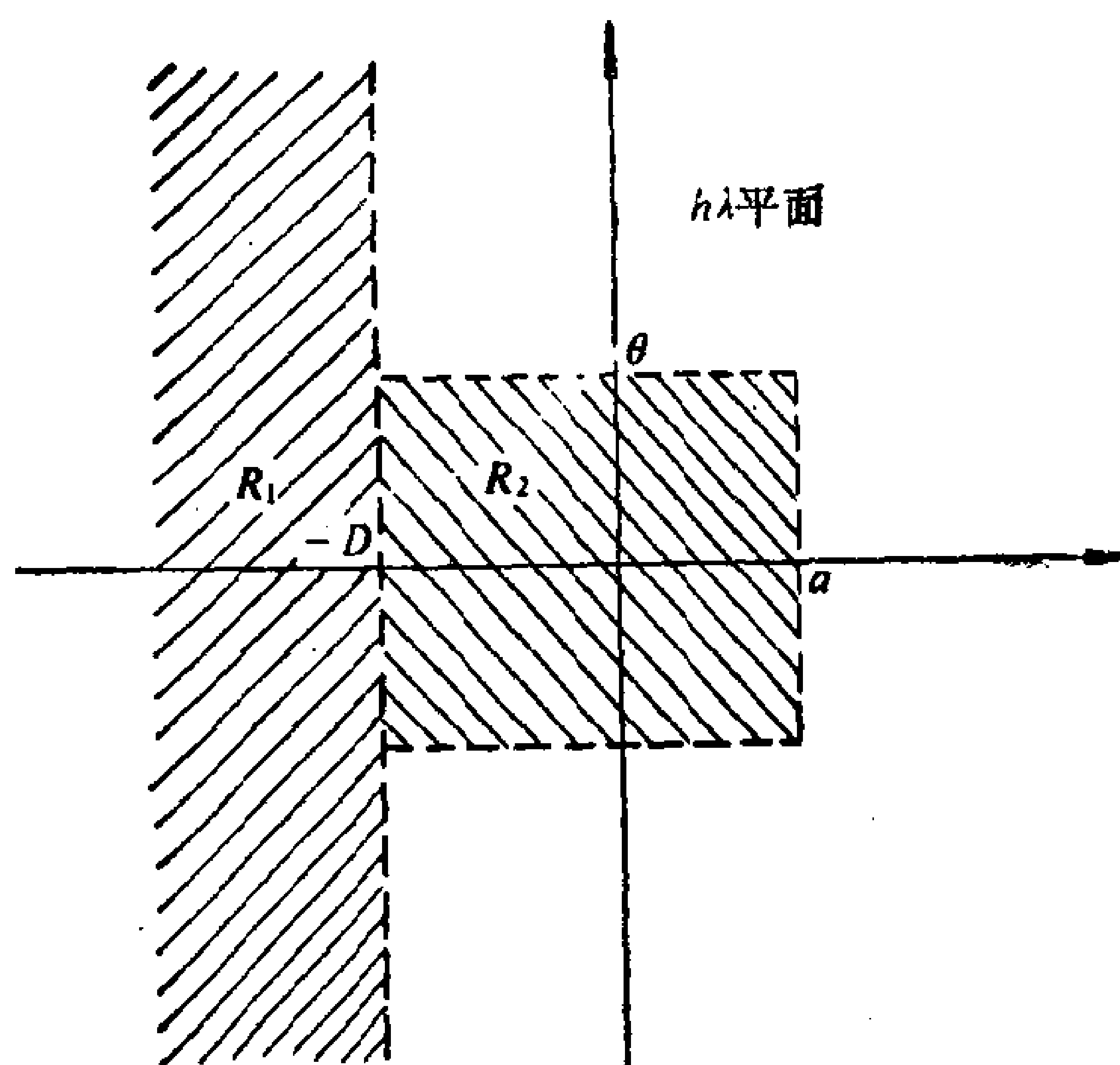


图 1.4　刚性稳定性

将 $h\lambda$ 限到区域 R_2 中．所以数值积分是不会在复平面的这部分上进行的．另外，如果 $|v|>\theta$，则积分一步，对应的量至少有 $\theta/2\pi$ 个完整的振荡周期．于是除了不考虑衰减量精度的区域 R_1 和不用的区域 $u>a$ 外，我们必须比较精确地描述这些振荡．为了保证数值精度，每个周期大约需要10个点，因此必须缩小步长使数值积分在 $|v|\leqslant\theta\leqslant\frac{\pi}{5}$ 的区域中进行．

上述表明，在实际求解刚性方程时，稳定性和精度对数值积分公式的要求在定义 1.7 中将能较好地反映出来．

有时 A 稳定性还不能完全反映我们对求解刚性方程的数值方法的稳定性要求．有一些方法虽然是 A 稳定的，稳定区域很大，但可能出现下面的现象：当 $h\lambda\to-\infty$ 时，有

$$|y_{n+1}/y_n|\to 1,$$

其中 $\{y_n\}$ 是该方法用定步长解方程(1.18)得到的序列．这种现象使得在解析解中一些非常快衰减到零的量在数值解中表现成缓慢地衰减，可能变为振荡的分量．例如用梯形法

$$y_{n+1} = y_n + \frac{h}{2}(y'_n + y'_{n+1})$$

求解方程(1.18)时,得到递推式

$$y_{n+1} = Q(h\lambda)y_n,$$

其中

$$Q(h\lambda) = \left(1 + \frac{h\lambda}{2}\right) \Big/ \left(1 - \frac{h\lambda}{2}\right).$$

对于任何有 $\mathrm{Re}\lambda < 0$ 的 λ 和任何固定的 $h > 0$, 都有 $|Q(h\lambda)| < 1$. 因此当 $n \to \infty$ 时, $y_n \to 0$, 所以梯形法是 A 稳定的. 但是

$$\left|\frac{y_{n+1}}{y_n}\right| = \left|\frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}\right|$$

$$= \left[\frac{1 + h\mathrm{Re}\,\lambda + \frac{1}{4}h^2|\lambda|^2}{1 - h\mathrm{Re}\,\lambda + \frac{1}{4}h^2|\lambda|^2}\right]^{\frac{1}{2}},$$

当 $|\mathrm{Re}\lambda| \gg 0$ 且 h 不太小时, $|y_{n+1}/y_n| \approx 1$. 因此对于大的 $\mathrm{Re}\lambda$, y_n 显示出衰减很慢的振荡. 这表示 A 稳定性对于保证方法的好的性质并不是充分的. 对于向后 Euler 公式

$$y_{n+1} = y_n + hy'_{n+1}$$

没有这种现象. 将其应用到(1.18)时,得到

$$y_{n+1} = \frac{1}{1 - h\lambda} y_n,$$

所以有

$$\left|\frac{y_{n+1}}{y_n}\right| = \left|\frac{1}{1 - h\lambda}\right|$$

$$= \left[\frac{1}{1 - 2h\mathrm{Re}\lambda + h^2|\lambda|^2}\right]^{\frac{1}{2}}.$$

因而可以得到对适当大的 h, 当 $\mathrm{Re}\lambda \to -\infty$ 时,有

$$\left|\frac{y_{n+1}}{y_n}\right| \to 0.$$

针对上述现象，提出了下述一些定义.

定义 1.8 数值积分公式称作是无限稳定的，如果下面的条件满足：存在实数 $w<0$，使得公式以定步长 $h>0$ 应用到方程(1.18)得到的序列 $\{y_n\}$ 满足

$$\sup_{\mathrm{Re}(h\lambda)<w} |y_{n+1}/y_n| = c < 1.$$

定义 1.9 数值积分公式称作是左稳定的(L稳定的)，如果它是A稳定的，并且以定步长 $h>0$ 应用到方程(1.18)得到的序列 $\{y_n\}$ 有递推式 $y_{n+1}=Q(h\lambda)y_n$，其中当 $\mathrm{Re}(h\lambda)\to-\infty$ 时有 $|Q(h\lambda)|\to 0$.

左稳定的名词是由 Ehle (1969) 引进的，Axelsson (1969)称其为刚性A稳定，而 Chipman(1971) 和 Axelsson (1972) 称其为强A稳定性.

在本书的后面可以看到，对于上述定义的每种稳定性，均可以构造具有这种稳定性的数值积分公式.

在文献中还定义了其它的许多稳定性，如B稳定性[34]，G稳定性[49]，S稳定性[95]等，本节所提出的几个稳定性的定义是最常用的. 在后面需要的地方再提出其它的稳定性的定义.

§3 一些刚性方程的例子

在这一节，我们从不同的应用领域中选取一些刚性方程的例子.

例 1.1 在化学反应中经常会遇到下面类型的反应速度方程

$$\begin{aligned}
y_1' &= -0.04y_1 + 10^4 y_2 y_3, & y_1(0) &= 1,\\
y_2' &= 0.04y_1 - 10^4 y_2 y_3 - 3\cdot 10^7 y_2^2, & y_2(0) &= 0, \quad (1.23)\\
y_3' &= 3\cdot 10^7 y_2^2, & y_3(0) &= 0,
\end{aligned}$$

其中 y_1，y_2，y_3 表示反应物的浓度. 方程组(1.23)的 Jacobi 矩阵

为

$$A=\begin{bmatrix}-0.04 & 10^4y_3 & 10^4y_2\\ 0.04 & -10^4y_3-6\cdot10^7y_2 & -10^4y_2\\ 0 & 6\cdot10^7y_2 & 0\end{bmatrix}.$$

容易证明,它是奇异的. 其三个特征值由

$$\lambda_1=0$$

和

$$\lambda^2+(0.04+10^4y_3+6\cdot10^7y_2)\lambda \\ +(0.24\cdot10^7y_2+6\cdot10^{11}y_2^2)=0$$

给出. 在 $t=0$ 时,三个特征值是 $\lambda_1=0$, $\lambda_2=0$, $\lambda_3=-0.04$. 这时还不是刚性的. 由(1.23)控制的反应过程将渐近地变到 $y_1=0$, $y_2=0$, $y_3=1$. 于是对于大的 t 值, A 的三个特征值将接近于 $\lambda_1=0$, $\lambda_2=0$, $\lambda_3=-10^4$. 事实上其特征值的变化如下

	λ_1	λ_2	λ_3
$t=0$	0	0	-0.04
$t=10^{-2}$	0	-0.36	-2180
$t=100$	0	-0.0048	-4240
$t=\infty$	0	0	-10^4

它的刚性比是很高的.

另外一个反应过程的微分方程组为

$$\begin{aligned}
y_1'&=y_3-100y_1y_2, & y_1(0)&=1,\\
y_2'&=y_3+2y_4-100y_1y_2-10^4y_2^2, & y_2(0)&=1,\\
y_3'&=-y_3+100y_1y_2, & y_3(0)&=0,\\
y_4'&=-y_4+10^4y_2^2, & y_4(0)&=0,
\end{aligned}\tag{1.24}$$

在开始时,它的 Jacobi 矩阵的特征值是-20101, -98.5, -2, 0, 然后它迅速地变成-200, $-1.1\pm2.5i$, 0, 以后慢慢地趋向于-177, $-1\pm0.6i$, 0, 所以(1.24)自始至终是刚性的.

例 1.2 在自动控制系统中经常会遇到如图 1.5 所示的闭合环节,其中 $W_1(P)$ 和 $W_2(P)$ 均是 P 的有理分式, P 为微分算子

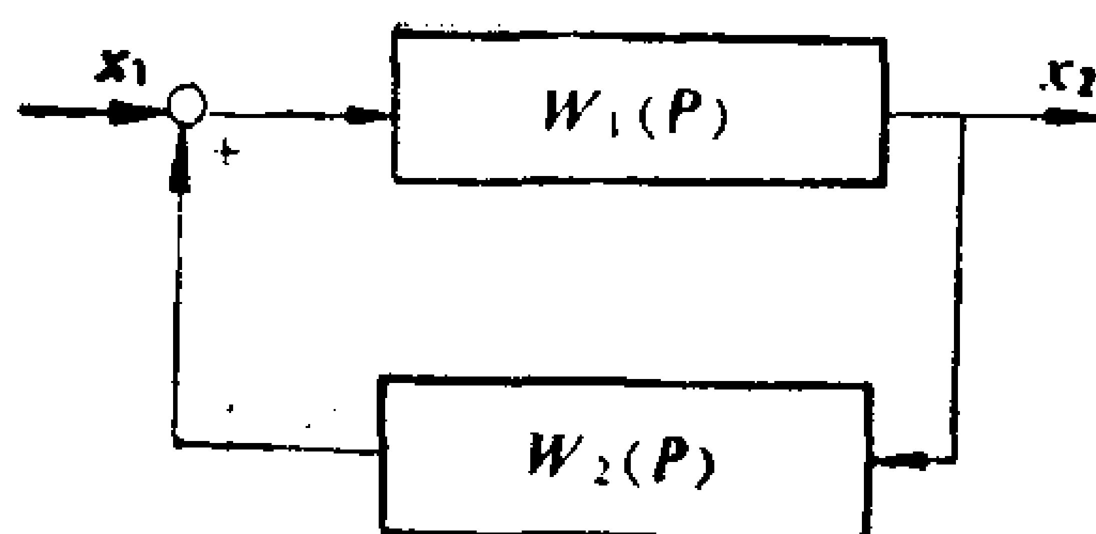


图 1.5

$\frac{d}{dt}$.

设 $W_1(P)$ 和 $W_2(P)$ 分别为

$$W_1(P)=\frac{K_1}{T_1P+1},\quad W_2(P)=\frac{K_2}{T_2P+1},$$

其中 T_1, T_2, K_1 和 K_2 均是给定的常数. 则输入 $x_1(t)$ 和输出 $x_2(t)$ 之间有关系式

$$x_2(t)=W_1(P)[x_1(t)+W_2(P)x_2(t)].$$

解出 $x_2(t)$, 得

$$x_2(t)=Q(P)x_1(t), \tag{1.25}$$

$Q(P)$ 是图 1.5 所示的环节的传递函数

$$Q(P)=\frac{W_1(P)}{1-W_1(P)W_2(P)}$$

$$=\frac{K_1(T_2P+1)}{(T_1P+1)(T_2P+1)-K_1K_2}.$$

(1.25)也可以写成二阶微分方程的形式

$$T_1T_2\frac{d^2x_2}{dt^2}+(T_1+T_2)\frac{dx_2}{dt}+(1-K_1K_2)x_2$$

$$=K_1T_2\frac{dx_1(t)}{dt}+K_1x_1(t), \tag{1.26}$$

这个方程的特征方程是

$$T_1T_2\lambda^2+(T_1+T_2)\lambda+(1-K_1K_2)=0,$$

当取

$$T_1 = 0.002,\ T_2 = \frac{1}{501},\ K_1 = 1,\ K_2 = \frac{499}{501}$$

时，其特征根为

$$\lambda_1 = -1000,\ \lambda_2 = -1.$$

因此，方程(1.26)是刚性方程，其刚性比是 1000.

例 1.3 描述宇航飞行器的运动轨道的常微分方程组是刚性方程. 为叙述简单起见，假定地球是不转的. 按下述的方式建立坐标系：坐标原点 o 在发射点，ox 轴在过发射点的地平面上，指向发射方向，oy 轴由 o 点垂直向上，oz 轴与 ox 和 oy 轴垂直，并且与 ox、oy 轴组成右手系. 假定飞行器在 oxy 平面内运动. 我们得到下面的经简化的运动方程

$$\begin{cases} mv' = P_v + C_v + X_v - mg\sin\theta, \\ mv\theta' = P_\theta + C_\theta + X_\theta - mg\cos\theta, \\ x' = v\cos\theta, \\ y' = v\sin\theta, \end{cases} \tag{1.27a}$$

$$\begin{cases} J\varphi'' + M_{阻} + M_{气} + M_{控} = 0, \\ \tau_2\delta'' + \tau_1\delta' + \delta = a_0(\varphi - \varphi_c), \end{cases} \tag{1.27b}$$

其中的记号有下面的物理意义：v 为飞行器的速度，P_v，C_v，X_v 分别是推力，控制力，气动力在速度方向上的投影，而 P_θ，C_θ，X_θ 分别是推力，控制力，气动力在与速度方向垂直向上的方向上的投影，g 是重力加速度，m 是飞行器在时刻 t 时的质量，θ 是速度方向与ox轴的夹角，称作速度倾角. φ 是飞行器的纵轴与 ox 轴的夹角，称作俯仰角. δ 是控制机构的偏角，称作舵偏角，用以产生控制力矩和控制力，使飞行器沿着规定的飞行轨道飞行. φ_c 是程序角，它规定 φ 在时刻 t 应取的量值，由它确定飞行器的飞行轨道. $\varphi - \varphi_c$ 即是 φ 与 φ_c 之间的偏差. a_0 是放大系数，J 是飞行器绕飞行器横轴的转动惯量. $M_{阻}$，$M_{气}$，$M_{控}$ 分别是阻尼力矩，气动力矩和控制力矩.

整个方程组(1.27)称作完全组. 它可以分成二个子组(1.27a)和(1.27b).(1.27a)描述飞行器的质心运动. 当飞行速度比较大

时，质心运动的惯性比较大，相对来说它的变化是慢的．(1.27*b*) 描述控制机构对飞行器的姿态的控制运动．由于运动的惯性较小，它是快变的．τ_2 和 τ_1 均是小参数，τ_1 的量级为 10^{-1}，而 τ_2 的量级为10^{-3}．通常积分(1.27*a*)不会遇到困难，其 Jacobi 矩阵的特征值的实部和虚部的绝对值很小．但是 (1.27*b*) 的性质不一样，它的 Jacobi 矩阵的特征值的实部的绝对值是很大的．例如我们只考虑 (1.27*b*) 的第二个方程．相应于这个方程的特征方程是

$$\tau_2\lambda^2 + \tau_1\lambda + 1 = 0,$$

其根为

$$\lambda_1 \approx -\frac{\tau_1}{\tau_2} + \frac{1}{\tau_1},$$

$$\lambda_2 \approx -\frac{1}{\tau_1}.$$

因此$|\lambda_2|$的量级为 10，而$|\lambda_1|$的量级为 10^2，所以精确地积分(1.27)需要较小的步长．

由上面的分析，完全组(1.27)是一个典型的刚性微分方程组，它包含分别含快变分量和慢变分量的二个微分方程组．

为了有效地处理(1.27) 的积分问题，在实践中采用下面二种方法．

(i) 应用奇异摄动的思想将方程组(1.27*b*)进行简化．这里又有二种方案．一种方案是只令 $\tau_2 = 0$，于是(1.27*b*)的第二个方程简化成

$$\tau_1\delta' + \delta = a_0(\varphi - \varphi_c), \tag{1.28}$$

可以降低(1.27)的刚性比．第二种方案是将(1.27*b*)简化成代数方程组

$$\begin{aligned} &M_{气} + M_{控} = 0, \\ &\delta = a_0(\varphi - \varphi_c). \end{aligned} \tag{1.29}$$

这二种方案得到的计算飞行器轨道的方程组均称作简化组，它们在工程实际中是普遍使用的．当(1.27)中参数的间断性小时，采用第二方案对计算的精度影响很小．若间断较大．可采用第一方

案，这样可保持一定的暂态过程，使得计算结果保持连续性.

(ii) 由于(1.27a)和(1.27b)具有不同的性质，并且具有一定的相对独立性，可以对它们使用不同的方法进行积分，详细情形见第十一章.

例 1.4　在一定的区域中用常微分方程组来近似偏微分方程是产生刚性方程的典型例子. 当将偏微分方程的空间导数用差分代替，就得到常微分方程组. 下面我们举一个求解非线性抛物型偏微分方程的例子. 考虑非线性抛物型偏微分方程的第一边值问题

$$
\begin{cases}
\dfrac{\partial u}{\partial t} = \dfrac{\partial}{\partial x}\left(F(u)\dfrac{\partial u}{\partial x}\right), \ x \in [0, 1], \ t \geqslant 0, \\
u(0, t) = 0 \\
u(1, t) = \sin t, \\
u(x, 0) = \begin{cases} 2x, \ 0 \leqslant x \leqslant \dfrac{1}{2}, \\ 2(1-x), \ \dfrac{1}{2} \leqslant x \leqslant 1, \end{cases}
\end{cases}
\tag{1.30}
$$

空间导数 $\dfrac{\partial u}{\partial x}$ 等用中心差分近似. 将区间 $[0,1]$ 分成$N+1$ 个长度为 Δx 的相等的子区间，即有 $\Delta x = \dfrac{1}{N+1}$. 令 $v(t)$ 是分量为

$$v_j(t) = u(j\Delta x, t), \ j = 0, 1, \cdots, N+1$$

的向量值函数，于是我们有初值问题

$$\frac{dv_j}{dt} = \frac{1}{(\Delta x)^2}[F(v_{j+\frac{1}{2}})(v_{j+1} - v_j) - F(v_{j-\frac{1}{2}})(v_j - v_{j-1})],$$

$$j = 1, \cdots, N,$$

$$v_j(0) = \begin{cases} 2(j\Delta x), \ 0 \leqslant j\Delta x \leqslant \dfrac{1}{2}, \\ 2(1 - j\Delta x), \ \dfrac{1}{2} \leqslant j\Delta x \leqslant 1. \end{cases}$$

其中 $v_0 = 0$, $v_{N+1} = \sin t$. 再取近似

$$F(v_{j\pm\frac{1}{2}}) = [F(v_{j\pm 1}) + F(v_j)]/2,$$

得到

$$\frac{dv}{dt} = \frac{1}{(\Delta x)^2}[Q(v)v + f(t)], \tag{1.31}$$

其中

$$v = (v_1, v_2, \cdots, v_N)^T, \ f(t) = (0, 0\cdots, \sin tF(v_{N+\frac{1}{2}}))^T.$$

$Q(v)$ 是三对角矩阵

$$Q(v) = \begin{bmatrix} q_{11} & q_{12} & 0 & & & \\ q_{21} & q_{22} & q_{23} & & 0 & \\ 0 & q_{32} & q_{33} & q_{34} & & \\ & \ddots & \ddots & \ddots & & \\ & & q_{jj-1} & q_{jj} & q_{jj+1} & \\ & 0 & \ddots & \ddots & \ddots & \\ & & & & & q_{N-1N} \\ & & & q_{NN-1} & q_{NN} & \end{bmatrix},$$

$$q_{jj-1} = \frac{1}{2}(F(v_j) + F(v_{j-1})),$$

$$q_{jj} = -\frac{1}{2}F(v_{j+1}) - F(v_j) - \frac{1}{2}F(v_{j-1}),$$

$$q_{jj+1} = \frac{1}{2}(F(v_j) + F(v_{j+1})).$$

方程(1.31)是非线性微分方程组，具有比较宽的时间常数带．对于 $F(u) \equiv 1$ 的情形，$\frac{1}{(\Delta x)^2}Q$ 的特征值为

$$\lambda_s = -4(N+1)^2\sin^2\frac{s\pi}{2(N+1)}, \ s = 1, 2, \cdots, N.$$

当N比较大时，最大的特征值为

$$\lambda_1 = -4(N+1)^2\sin^2\frac{\pi}{2(N+1)},$$

$$\approx -4(N+1)^2\left(\frac{\pi}{2(N+1)}\right)^2 = -\pi^2,$$

而最小的特征值为

$$\lambda_N = -4(N+1)^2\left(\sin \frac{N\pi}{2(N+1)}\right)^2 \approx -4(N+1)^2.$$

N 愈大，方程组(1.31)的刚性比愈大.

例 5　考虑无约束极小化问题

$$\min_{x \in D} f(x), \tag{1.32}$$

其中 $f: D \subset R^n \to R^1$ 是在 D 上为连续 Fréchet 可微的. D 是开凸集，包含满足 $\nabla f(x^*) = 0$ 的点 x^*. 令 f 在 x^* 处二次 Fréchet 可微，并且假定 $f''(x^*)$ 是正定的. 这表示 x^* 是 f 的局部无约束极小值点. 在无约束极小化问题(1.32)中，二阶导数 $f''(x^*)$ 的条件数可能很大，这时问题称作病态的.

病态的问题往往是由于各个变量的比例不同引起的，特别是当 x 的维数大时更易产生. 有些病态的问题是用逐次无约束极小化方法来求具有约束的问题时产生的. 例如求具有等式约束的极小化问题

$$\min_{g(x)=0} F(x), \tag{1.33}$$

选取足够大的 P，作罚函数 $f(x) = F(x) + P\|g(x)\|^2$. 则(1.33)的求解可近似地转化成无约束极小化问题(1.32)的求解. 但当 P 很大时，这时的问题是病态的.

当用梯度法求解问题(1.32)时，得到

$$x^{k+1} = x^k - r^k \nabla f(x^k),\ x^0 \text{ 给定}, \tag{1.34}$$

其中 x^k 是第 k 次迭代值，而 r^k 是正步长. (1.34)可以看成是积分微分方程组

$$\frac{dx}{dt} = -\nabla f(x),\ x(0) = x^0 \tag{1.35}$$

的 Euler 方法. 为了求解问题(1.32)，只要求出初值问题(1.35)的解. 当 x 接近 x^* 时，将(1.35)右边 Taylor 展开，得到

$$\frac{dx}{dt} = -f''(x^*)(x - x^*) + R(x), \tag{1.36}$$

其中 $R(x)$ 是 $\|x - x^*\|$ 的二阶项. 由于 $f''(x^*)$ 的条件数大，

(1.36)是刚性常微分方程.

§4 稳定区域的计算

现在讨论§1中定义1.3所定义的稳定区域的求法. 因为数值积分公式的稳定区域的图形可以帮助理解这个公式的性能.

对于单步法,若将它用于试验方程 $y' = \lambda y$, 一般可以得到两步间的关系的如下形式

$$y_{n+1} = \phi(h\lambda)y_n.$$

因此,令 $z = h\lambda$,

$$|\phi(z)| < 1$$

是方法的稳定条件. z 复平面上满足稳定条件的 z 形成的区域是这个方法的稳定区域.

例如 Euler 法的稳定条件是(1.20), 其稳定区域是以 $-1+0i$ 为中心以1为半径的圆的内部,如图1.1所示.

将向后 Euler 公式

$$y_{n+1} = y_n + hf_{n+1},\quad f_{n+1} = f(t_{n+1}, y_{n+1}),$$

用于试验方程,得

$$y_{n+1} = (1 - h\lambda)^{-1}y_n,$$

稳定条件是

$$|(1 - z)^{-1}| < 1,$$

稳定区域是以 $1 + 0i$ 为中心以1为半径的圆的外部.

将梯形法

$$y_{n+1} = y_n + \frac{1}{2}h(f_n + f_{n+1})$$

用于试验方程,得

$$y_{n+1} = \frac{1 + \frac{1}{2}h\lambda}{1 - \frac{1}{2}h\lambda}y_n$$

的稳定条件是

$$\left|\frac{1+\frac{1}{2}z}{1-\frac{1}{2}z}\right|<1.$$

设 $z=\alpha+i\beta$，稳定条件化为

$$\left(1+\frac{1}{2}\alpha\right)^2+\frac{1}{4}\beta^2<\left(1-\frac{1}{2}\alpha\right)^2+\frac{1}{4}\beta^2,$$

当且仅当 $\alpha=\mathrm{Re}z<0$ 时，上式成立. 所以梯形法的稳定区域是不包括虚轴的整个左半平面.

对于比较复杂的单步法，用分析的方法不能完全定出稳定区域时，可以借助于数值计算.

例如四阶显式 Runge-Kutta 公式的稳定条件是

$$|\phi(z)|<1.$$

由(1.22)知

$$\phi(z)=1+z+\frac{1}{2}z^2+\frac{1}{6}z^3+\frac{1}{24}z^4.$$

因为当 $|z|\to+\infty$ 时 $|\phi(z)|\to+\infty$，所以稳定区域是有界的. $|\phi(z)|=1$ 是封闭曲线，我们叫它做曲线 C. 一般，曲线 C 将 z 平面分成若干连通的区域，判断那个区域属于稳定区域，把这些区域合在一起就是稳定区域. 实际上只要在区域内检查一个点是否满足稳定条件就可以代表这个区域是否属于稳定区域，这个事实用 $\phi(z)$ 的连续性可以证明.

我们讨论如何用数值方法求出曲线 C. 令

$$\phi(z)=e^{i\theta}, \tag{1.37}$$

取 $0\leqslant\theta\leqslant2\pi$，解出对应于每个 θ 值的(1.37)的根，根的全体就是曲线 C. 由于 $\phi(\bar{z})=\overline{\phi(z)}$，所以只要取 $0<\theta\leqslant\pi$，再利用共轭性质就可以得到全部 C 的点.

实际计算时，取满足 $0\leqslant\theta\leqslant\pi$ 的一系列离散点，例如取 $\theta_j=j\pi/N$，$j=0,1,\cdots,N$，对于每个 θ_j 解方程(1.37)，求出曲

线 C 上的四个点．例如可以用下山法[16]求(1.37)的复根．若取得 N 够大，联接求出的根便得到曲线 C 的近似图形．

另一种方法是用根轨迹的思想[14]求曲线 C．将方程的根 z 看做 θ 的函数，将方程(1.37)两端对 θ 微分，得

$$\psi'(z)\frac{dz}{d\theta}=ie^{i\theta},$$

也得到曲线 C 的微分方程

$$\frac{dz}{d\theta}=i\psi(z)/\psi'(z), \tag{1.38}$$

若知道一个初值，例如 $\theta=0$ 时 $z=1$，由数值积分微分方程(1.38)就可以求出曲线 C．

对于一般情形，只要 $\psi'(z)\neq 0$，数值积分就可进行．如果积分到某一步得到 z 后，$\psi'(z)=0$．在点 z 上可以用 Taylor 展开的方法，直接求对应于 $\Delta\theta$ 的 Δz．设在点 z 上

$$\psi'(z)=\cdots=\psi^{(k-1)}(z)=0,\psi^{(k)}(z)\neq 0,$$

将 $\theta+\Delta\theta$ 和 $z+\Delta z$ 代入方程(1.37)和原方程相减，得

$$\frac{1}{k!}\psi^{(k)}(z)\Delta z^k+\frac{1}{(k+1)!}\psi^{(k+1)}(z)\Delta z^{k+1}+\cdots$$
$$=e^{i\theta}(e^{i\Delta\theta}-1),$$

略去左端第二项及以后的项，可以解出 k 个 Δz 的值．若认为这样求出的 Δz 不够准确，可以利用某种迭代使 Δz 精确化．例如用迭代式

$$\Delta_z{}^{(m+1)}=\Big[e^{i\theta}(e^{i\Delta\theta}-1)\Big/\Big(\frac{1}{k!}\psi^{(k)}(z)$$
$$+\frac{1}{(k+1)!}\psi^{(k+1)}(z)\Delta z^{(m)}+\cdots\Big)\Big]^{1/k},$$

当 $\Delta z\to 0$ 时，迭代式的收敛条件是

$$\left|\frac{\psi^{(k+1)}(z)}{k(k+1)\psi^{(k)}(z)}\right|\cdot\left|\frac{k!}{\psi^{(k)}(z)}(e^{i\Delta\theta}-1)\right|^{1/k}<1,$$

只要 $\Delta\theta$ 取得足够小，这个条件就可满足．

若稳定区域的形状简单，如四阶显式 Runge-Kutta 公式的稳定区域(见图 1.2)，也可以直接从定义计算稳定区域的图形，仍以四阶显式 Runge-Kutta 公式为例，可将坐标原点移到 $-1+0i$，做一系列过新原点的射线，在射线上取一些点直接计算 $|\phi(z)|$，若 $|\phi(z)|<1$ 则这个点属于稳定区域，否则不属于稳定区域．求出 $|\phi(z)|=1$，即属于曲线 C 的点，即可逐步求出稳定区域的边界曲线 C．

单步法的稳定区域有如下性质：

1. 由于 $\phi(\bar{z})=\overline{\phi(z)}$，稳定区域对实轴对称．

2. 只要方法是收敛的，$\phi(z)$ 在 $z=0$ 应是 e^z 的近似式，因而有

$$\phi(z)=1+z+O(|z|^2),$$

由此可知，原点属于曲线 C，在实轴上原点的邻域左边属于稳定区域，右边不属于稳定区域．

3. 可以证明 $\phi'(0)=1$，又由 $\phi(0)=1$，由(1.38)知

$$\left.\frac{dz}{d\theta}\right|_{\theta=0}=i$$

即曲线 C 在原点与虚轴相切．

对于四阶显式 Runge-Kutta 公式，可以看出

$$\phi^{(k)}(0)=1,\ k=0,1,2,3,4,\ \phi^{(5)}=\phi^{(6)}=0.$$

并能证明

$$\left.\frac{dz}{d\theta}\right|_{\theta=0}=i,\ \left.\frac{d^2z}{d\theta^2}\right|_{\theta=0}=\left.\frac{d^3z}{d\theta^3}\right|_{\theta=0}=\left.\frac{d^4z}{d\theta^4}\right|_{\theta=0}=0,$$

$$\left.\frac{d^5z}{d\theta^5}\right|_{\theta=1}=i,\ \frac{d^6z}{d\theta^6}=4.$$

由此可知，曲线 C 在原点与虚轴相切，离开原点后经过第一象限．并能算出在点 $0+2\sqrt{2}\,i$ 通过虚轴．还能近似地算出曲线 C 在点 $-2.785+0i$ 通过负实轴．

我们现在讨论线性多步法稳定区域的求法．设线性 k 步法为

$$\rho(E)y_n=h\sigma(E)f_n,$$

这里

$$\rho(s)=\sum_{i=0}^{k}\alpha_i s^{k-i},\qquad \sigma(s)=\sum_{i=0}^{k}\beta_i s^{k-i},$$

$$Ey_n=y_{n+1},\ \alpha_0\neq 0,\ k>1.$$

将它应用于试验方程 $y'=\lambda y$，并令 $y_{n+1}=\mu y_n$，则有

$$\rho(\mu)=h\lambda\sigma(\mu). \tag{1.39}$$

给定 $z=h\lambda$，由(1.39)可以解出 k 个根 $\mu_1,\cdots,\mu_k$. μ 是 z 的多值函数. 方法的稳定区域是使 $\max\limits_{1\leqslant j\leqslant k}|\mu_j|<1$ 的 z 的区域. 和上述的讨论相似，我们先求出使某个 $|\mu_j|=1$ 的 z 所形成的曲线 C，即

$$C=\{z\,|\,|\mu_1|=1\text{ 或}|\mu_2|=1\cdots\text{ 或}|\mu_k|=1\}.$$

求曲线 C 时不必解方程(1.39). 由于(1.39)的特殊形式，可以给出所有使 $|\mu|=1$ 的 μ 的值，算出的 z 值就形成曲线 C. 即是由(1.39)解出 $z=h\lambda$，也即

$$z=\rho(\mu)/\sigma(\mu),$$

令 $\mu=e^{i\theta}$，$0\leqslant\theta\leqslant\pi$，便可求出曲线 C. 具体计算时给 θ 一系列离散值，如令 $\theta_j=j\pi/N, j=1,\cdots,N$，便可算出 z 的一系列离散值，用这些值可以描出 C 的近似曲线.

曲线 C 将 z 复平面分成若干连通的区域. 在每个区域内任选一个 z 值，考查它所确定的 $\mu_j(j=1,\cdots,k)$ 是否满足

$$\max_{1\leqslant j\leqslant k}|\mu_j|<1,$$

若满足则这个区域属于稳定区域，否则不属于稳定区域. 这是因为方程(1.39)的连续性和在每个区域内满足 $|\mu_j|<1$ 的 μ_j 的个数不变. 反之，若在一个区域内有两点 z_1 和 z_2，对应的 $|\mu_j|<1$ 的 μ_j 的个数不同. 在这个区域内用一条曲线联结 z_1 和 z_2，取 z 由 z_1 沿此曲线连续地变到 z_2 时，对应 z_1 的 k 个 μ_j 连续地变到对应 z_2 的 k 个 μ_j，这时必定有一 μ_j，在某个点上使 $|\mu_j|=1$，但这个点应该在曲线 C 上，这与联结 z_1 和 z_2 的曲线在区域内部相

矛盾．这个矛盾证明了在一个区域内 $|\mu_i|<1$ 的 μ_i 的个数不变．

在文章[19]中，用这个方法算出了四阶以内的 Adams 公式（内插型和外插型）的稳定区域。

第二章　线性多步公式的稳定性

多步方法是利用前面几步得到的关于常微分方程解的信息来构造解在新的节点上的近似值．本章介绍求解刚性方程的线性多步方法的稳定性理论．与其它方法相比，它的理论比较成熟，并且形成的概念已应用来研究许多其它方法的稳定性．

§1　线性多步公式

考虑一阶常微分方程组的初值问题

$$\frac{dy}{dt}=f(t,y),\quad y(0)=y_0, \tag{2.1}$$

其中 y, f 均是m维向量，$t\geqslant 0$．设 $y(t)$ 为它的精确解．令 $t_n=nh$, $n=0,1,\cdots$, y_n 为 $y(t)$在 t_n 处的近似值．计算 y_n 数值的一般常系数的线性 k 步公式（简称 k 步公式）为

$$\begin{aligned}&\alpha_k y_{n+k}+\alpha_{k-1}y_{n+k-1}+\cdots+\alpha_0 y_n\\&\quad=h(\beta_k f_{n+k}+\beta_{k-1}f_{n+k-1}+\cdots+\beta_0 f_n).\end{aligned} \tag{2.2}$$

假定α_i, β_i, $i=0,1,2,\cdots,k$ 是实常数，$\alpha_k\neq 0$．h 是正常数，称为步长，$f_n=f(t_n,y_n)$．如果给定计算的起始向量 y_0, y_1, $\cdots$, y_{k-1}，则可由公式(2.2)递推地计算 y_k, y_{k+1}, $\cdots$．当 $\beta_k=0$ 时，公式称为显式公式，否则称为隐式公式．关于线性多步公式的详细理论可在 Henrici [62] 中找到，这一节只将本书中一些常用的名词和结果列出．

对差分方程(2.2)，引进多项式

$$\rho(\zeta)=\sum_{i=0}^{k}\alpha_i\zeta^i,\qquad \sigma(\zeta)=\sum_{i=0}^{k}\beta_i\zeta^i \tag{2.3}$$

和算子

$$L = \rho(E) - hD\sigma(E), \tag{2.4}$$

其中 $D = d/dt$，E 是位移算子，定义成

$$Ey(t) = y(t+h), \text{ 或 } Ey_n = y_{n+1}. \tag{2.5}$$

假定 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 无公因子．对于充分光滑的任意函数 $\phi(t)$，将 $L\phi(t)$ 中出现的 $\phi(t+jh)$ 及其导数 $\phi'(t+jh)$ 在 t 处展成 Taylor 级数，则有

$$L\phi(t) = c_0\phi(t) + c_1h\phi'(t) + \cdots + c_sh^s\phi^{(s)}(t) + \cdots. \tag{2.6}$$

如果在(2.6)中 $c_0 = c_1 = \cdots = c_p = 0$，而 $c_{p+1} \neq 0$，则称公式(2.2) 的阶为 p．于是对于任意的 $p+1$ 次连续可微的函数 $\phi(t)$，有

$$L\phi(t) \sim c_{p+1}h^{p+1}\phi^{(p+1)}(t)(h \to 0), \tag{2.7}$$

上式表示 $L\phi(t)$ 具有 $O(h^{p+1})$．数 c_{p+1} 和 p 与函数 $\phi(t)$ 是无关的，只依赖于 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 的系数．当 $p \geqslant 1$ 时，公式(2.2)称作相容的．容易证明，公式(2.2)的相容性由关系式

$$\rho(1) = 0, \ \rho'(1) = \sigma(1) \tag{2.8}$$

来表示． 并且有 $\sigma(1) \neq 0$，因为否则 ρ 和 σ 将具有公因子． 量 $c^* = -c_{p+1}/\sigma(1)$ 称作公式(2.2)的误差常数． 对于具有阶为 p 的方法的精度，它是比较的一个适当的度量．

通过选取 (2.7)中的一个适当的特殊的 $\phi(t)$，可以确定常数 c_{p+1} (或 c^*) 和 p．取 $\phi(t) = e^t$，并令 $e^h = \zeta$，于是

$$\rho(\zeta) - \sigma(\zeta)\log\zeta \sim c_{p+1}(\zeta-1)^{p+1} \quad (\zeta \to 1),$$

因此有

$$\log\zeta - \rho(\zeta)/\sigma(\zeta) \sim c^* \cdot (\zeta-1)^{p+1} \ (\zeta \to 1). \tag{2.9}$$

对于相容的方法，多项式 $\rho(\zeta)$ 有根 $+1$，称这个根为公式(2.2)的主根，将其表成 ζ_1，其余的根 ζ_s，$s = 2, \cdots, k$ 称作寄生根．这是由于方法的步数大于 1 引起的．为了使公式(2.2)能实际用来求解(2.1)，必须仔细控制这些寄生根的位置．通常要求 $\rho(\zeta)$ 的根的模均不超过 1，并且模为 1 的根均是单根． 满足这种条件的 k 步公式(2.2)称作是零稳定的．简称为稳定的．

Dahlquist 关于线性多步公式的一个基本结果为：线性多步公

式(2.2)收敛的充分必要条件为它是相容的和稳定的.

一般的 k 步公式 (2.2) 具有 $2k+2$ 个系数. Dahlquist (1956)[47] 证明可以选取多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 使公式(2.2)的阶 $p=2k$, 但不能达到比 $2k$ 更高的阶. 这个结果没有实际价值,因为这些高阶方法不满足零稳定性条件. Dahlquist 证明零稳定的 k 步公式(2.2)的阶不能超过 $k+1$ (如果 k 是奇数)或 $k+2$ (如果 k 是偶数).

考虑试验方程

$$\frac{dy}{dt}=\lambda y,\quad \mathrm{Re}\,\lambda<0,$$

当公式(2.2)以步长 h 应用到这个方程时, (2.2) 变成常系数的差分方程,其特征方程为

$$\rho(\zeta)-q\sigma(\zeta)=0, \tag{2.10}$$

$q=h\lambda$. 设(2.10)的根为 $\zeta_i(q)$, $i=1,\cdots,k$, 于是定义 1.3 中的稳定区域可以表成

$$R=\{q=h\lambda\,|\,|\zeta_i(q)|<1,\ i=1,\cdots,k\},$$

称它为绝对稳定区域. 我们还可以定义相对稳定区域. 令 $\zeta_1(q)$ 为对应于主根 ζ_1 的一支根,则相对稳定区域为

$$R_r=\{q\in\Omega\,|\,|\zeta_i(q)|<|\zeta_1(q)|,\ i=2,3,\cdots,k\}$$

Ω 是 $\zeta_i(q)$ 可解析延拓的最大的星形区域.

§2 线性多步公式的 A 稳定性

Dahlquist 在[48]中引入 A 稳定性的定义后, 立即提出下面的结果.

定理 2.1 k 步公式 (2.2) 是 A 稳定的充分必要条件为对于 $|\zeta|>1$, $\rho(\zeta)/\sigma(\zeta)$ 是正则的,并且具有非负的实部.

证明 将公式(2.2)应用到试验方程,(2.2)变成具有常系数的差分方程,其特征方程为(2.10).

必要性 设公式(2.2)是 A 稳定的,即对所有 $h\lambda$, 若 $\mathrm{Re}(h\lambda)<$

0,有(2.10)的所有根 ζ 均有 $|\zeta| < 1$. 所以对于任何 ζ,若 $|\zeta| > 1$ 并且是(2.10)的根,则相应的 $\mathrm{Re}(h\lambda) = \mathrm{Re}\{\rho(\zeta)/\sigma(\zeta)\} \geqslant 0$.

又设 ζ_1 是 $\sigma(\zeta) = 0$ 的根,且 $|\zeta_1| > 1$. 据 $\rho(\zeta)$ 与 $\sigma(\zeta)$无公因式的假定, $\rho(\zeta_1) \neq 0$, 故在 ζ_1 的某个领域中,存在正整数 m,使得有

$$\rho(\zeta)/\sigma(\zeta) \sim a(\zeta - \zeta_1)^{-m}, \quad a \neq 0,$$

显然在这个邻域中存在 ζ, 使 $\mathrm{Re}\{\rho(\zeta)/\sigma(\zeta)\} < 0$ 成立, 这与 $|\zeta| > 1$, $\mathrm{Re}\{\rho(\zeta)/\sigma(\zeta)\} \geqslant 0$ 矛盾. 因此如果 $|\zeta| > 1$, 有 $\sigma(\zeta) \neq 0$. 即 $\rho(\zeta)/\sigma(\zeta)$ 对于 $|\zeta| > 1$ 是正则的.

充分性 设 k 步公式 (2.2) 对 $|\zeta| > 1$, $\rho(\zeta)/\sigma(\zeta)$ 正则且 $\mathrm{Re}\{\rho(\zeta)/\sigma(\zeta)\} \geqslant 0$. 由后者可知,若 $\mathrm{Re}(h\lambda) < 0$ 则(2.10)的相应的根 ζ 均有 $|\zeta| \leqslant 1$, 并且这 ζ 不能有 $\sigma(\zeta) = 0$,否则 ζ 将是 σ 与 ρ 的公共根. 因此 $\rho(\zeta)/\sigma(\zeta)$ 在(2.10)的根处不奇异. 由此推得若 $\mathrm{Re}(h\lambda) < 0$, 则(2.10) 的根 ζ 不能有 $|\zeta| = 1$, 而一定有 $|\zeta| < 1$,这即是 A 稳定性.

推论 2.1 k 步公式是 A 稳定的,则由

$$u(\zeta) = \mathrm{Re}\{\rho(\zeta)/\sigma(\zeta)\} \tag{2.11}$$

定义的函数 $u(\zeta)$ 对所有实值 Θ 成立

$$u(e^{i\Theta}) \geqslant 0. \tag{2.12}$$

推论 2.2 如果 k 步公式是 A 稳定的,则多项式 $\sigma(\zeta)$ 的根 σ_i, $i = 1, \cdots, k$ 满足 $|\sigma_i| \leqslant 1$

利用与定理 1 类似的讨论给出下面的结果.

定理 2.2 显式的 k 步公式不能是 A 稳定的

证明 设 $\beta_k = 0$, 于是对某个整数 $m \geqslant 1$, 当 $\zeta \to \infty$ 时,有 $\sigma(\zeta) \sim a\zeta^{k-m}$, $a \neq 0$. 但是 $\rho(\zeta) \sim \alpha_k\zeta^k$, $\alpha_k \neq 0$. 因此有 $\rho(\zeta)/\sigma(\zeta) \sim b\zeta^m$, $\zeta \to \infty$, 其中 $b \neq 0$, $m \geqslant 1$, 这与对单位圆外的所有 ζ 具有非负的实部是矛盾的.

下面的例子说明相容的 A 稳定多步公式是存在的.

例 2.1 下面的线性多步公式均是 A 稳定的.

(i) 向后 Euler 公式

$$y_{n+1} = y_n + hf_{n+1}$$

对应的多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 为

$$\rho(\zeta) = \zeta - 1, \quad \sigma(\zeta) = \zeta.$$

(ii) 梯形法

$$y_{n+1} = y_n + \frac{h}{2}[f_{n+1} + f_n],$$

$$\rho(\zeta) = \zeta - 1. \quad \sigma(\zeta) = \frac{1}{2}(\zeta + 1).$$

(iii) $$y_{n+k} = y_n + \frac{1}{2}kh[f_{n+k} + f_n],$$

$$\rho(\zeta) = \zeta^k - 1, \quad \sigma(\zeta) = \frac{1}{2}k(\zeta^k + 1).$$

这些例子说明对任何正整数 k，存在相容的 A 稳定 k 步公式.

为了验证定理 1 给出的 A 稳定性的条件，需要在二维集合 $|\zeta| > 1$ 上判定 $q(\zeta) = \rho(\zeta)/\sigma(\zeta)$ 是否具有非负的实部. 下面给出 A 稳定性的一个充分条件，它仅需在一维区域上进行验证.

设多项式 $\sigma(\zeta)$ 的根 σ_i 满足条件

N_1 $$|\sigma_i| < 1, \; i = 1, 2, \cdots, k, \tag{2.13}$$

于是 $\sigma(e^{i\Theta}) \neq 0$. 存在数 δ_1 和 δ_2 使不等式

$$0 < \delta_1 \leqslant |\sigma(e^{i\Theta})| \leqslant \delta_2$$

成立. 条件

$$u(e^{i\Theta}) = |\sigma(e^{i\Theta})|^{-2}\mathrm{Re}\{\rho(e^{i\Theta})\sigma(e^{-i\Theta})\} \geqslant 0$$

等价于条件

$$\mathrm{Re}\,[\rho(e^{i\Theta})\sigma(e^{-i\Theta})] = \sum_{j=0}^{k} \gamma_j \cos j\Theta \geqslant 0, \tag{2.14}$$

其中

$$\gamma_0 = \sum_{l=0}^{k} \alpha_l\beta_l$$

$$\gamma_j = \sum_{l=0}^{k-j} (\alpha_{l+j}\beta_l + \alpha_l\beta_{l+j}), \quad 1 \leqslant j \leqslant k,$$

熟知，$\cos j\Theta = T_j(\xi)$，其中 $\xi = \cos\Theta$，而 $T_j(\xi)$ 是 Чебышев 多项式. 前 5 个多项式为

$$T_0(\xi) = 1,$$
$$T_1(\xi) = \xi,$$
$$T_2(\xi) = 2\xi^2 - 1,$$
$$T_3(\xi) = 4\xi^3 - 3\xi,$$
$$T_4(\xi) = 8\xi^4 - 8\xi^2 + 1,$$

一般有递推式

$$T_j(\xi) = 2\xi T_{j-1}(\xi) - T_{j-2}(\xi), \quad j = 2, 3, \cdots.$$

因此，如果条件(2.13)成立，则条件(2.12)等价于

N_2 $$P_k(\xi) \geqslant 0, \quad -1 \leqslant \xi \leqslant 1, \tag{2.15}$$

其中

$$P_k(\xi) = \sum_{j=0}^{k} \gamma_j T_j(\xi). \tag{2.16}$$

定理 2.3 如果条件(2.13)和(2.15)成立，则 k 步公式是 A 稳定的.

证明 设条件(2.13)成立，则在区域 $|\zeta| \geqslant 1$ 上 $q(\zeta)$ 是解析的，而 $u(\zeta)$ 是调和的. 由极小原理，对于 $|\zeta| \geqslant 1$ 有

$$u(\zeta) \geqslant \min_{\Theta} u(e^{i\Theta}).$$

如果条件(2.15)成立，则 $\min\limits_{\Theta} u(e^{i\Theta}) \geqslant 0$，于是由定理 1 的充分性证明推得 k 步公式是 A 稳定的.

充分条件(2.13)和(2.15)是容易验证的. 条件(2.15)表示只须在区间 $-1 \leqslant \xi \leqslant 1$ 上验证 $P_k(\xi) \geqslant 0$，而不需在二维集合 $|\zeta| > 1$ 上验证 $u(\zeta) \geqslant 0$. 为验证条件(2.13)，可通过变换

$$z = (\zeta + 1)/(\zeta - 1), \zeta = (z + 1)/(z - 1) = \zeta(z). \tag{2.17}$$

并令

$$r(z) = \left(\frac{z-1}{2}\right)^k \rho(\zeta(z)) = \sum_{j=0}^{k} a_j z^j, \tag{2.18}$$

$$s(z)=\left(\frac{z-1}{2}\right)^k\sigma(\zeta(z))=\sum_{j=0}^{k}b_jz^j, \tag{2.19}$$

则多项式 $\sigma(\zeta)$ 的根 σ_i 满足条件(2.13)的充要条件是多项式 $s(z)$ 的根 $s_i, i=1,\cdots,k$ 满足 $\mathrm{Re}s_i<0$. 这是因为变换 $z=(\zeta+1)/(\zeta-1)$ 将 ζ 平面的单位圆 $|\zeta|<1$ 变换成 z 平面的开左半平面，圆周 $|\zeta|=1$ 变成 z 平面上的虚轴. 由 Routh 准则很易验证 $\mathrm{Re}s_i<0$ 是否成立.

例 2.2 考虑线性二步公式族

$$\begin{aligned}\frac{1}{2}(1-3\beta_0-\beta_1)y_n-2(1-\beta_0)y_{n+1}\\ +\frac{1}{2}(3-\beta_0+\beta_1)y_{n+2}\\ =h(\beta_0f_n+\beta_1f_{n+1}+f_{n+2}),\end{aligned} \tag{2.20}$$

其中 β_0, β_1 是自由参数，这个族中的公式是至少具有二阶精度 ($p\geqslant 2$) 的线性二步公式.

现在来确定当参数 β_0, β_1 在什么条件下,(2.20)为 A 稳定. 通过变换(2.17),相应于(2.20)的多项式 $\sigma(\zeta)=\zeta^2+\beta_1\zeta+\beta_0$ 变换成多项式

$$s(z)=\frac{1}{4}(\bar{b}_2z^2+\bar{b}_1z+\bar{b}_0),$$

其中

$$\begin{aligned}\bar{b}_0&=1+\beta_0-\beta_1,\\ \bar{b}_1&=2(1-\beta_0),\\ \bar{b}_2&=1+\beta_0+\beta_1.\end{aligned}$$

由 Routh 准则，为了使 $s(z)$ 的零点均有负的实部必须要求 $\bar{b}_0$, $\bar{b}_1$, $\bar{b}_2$ 均不为零,并具有同一符号. 容易看出,这三个量不可能同时是负的,因此条件(2.13)等价于

$$\begin{aligned}1-\beta_0&>0,\\ 1+\beta_0+\beta_1&>0,\\ 1+\beta_0-\beta_1&>0.\end{aligned} \tag{2.21}$$

满足不等式（2.21）的所有（β_0, β_1）组成图 2.1 中的开三角形阴影区．对公式(2.20)考察条件(2.15)，得到

$$\gamma_0 = \frac{3}{2}\gamma(\beta_0, \beta_1),$$

$$\gamma_1 = -2\gamma(\beta_0, \beta_1),$$

$$\gamma_2 = \frac{1}{2}\gamma(\beta_0, \beta_1)$$

和

$$P_2(\xi) = \frac{1}{2}\gamma(\beta_0, \beta_1)[3T_0(\xi) - 4T_1(\xi) + T_2(\xi)]$$

$$= (1-\xi)^2\gamma(\beta_0,\beta_1),$$

其中

$$\gamma(\beta_0,\beta_1) = 1 - \beta_1 + \beta_0\beta_1 - \beta_0^2.$$

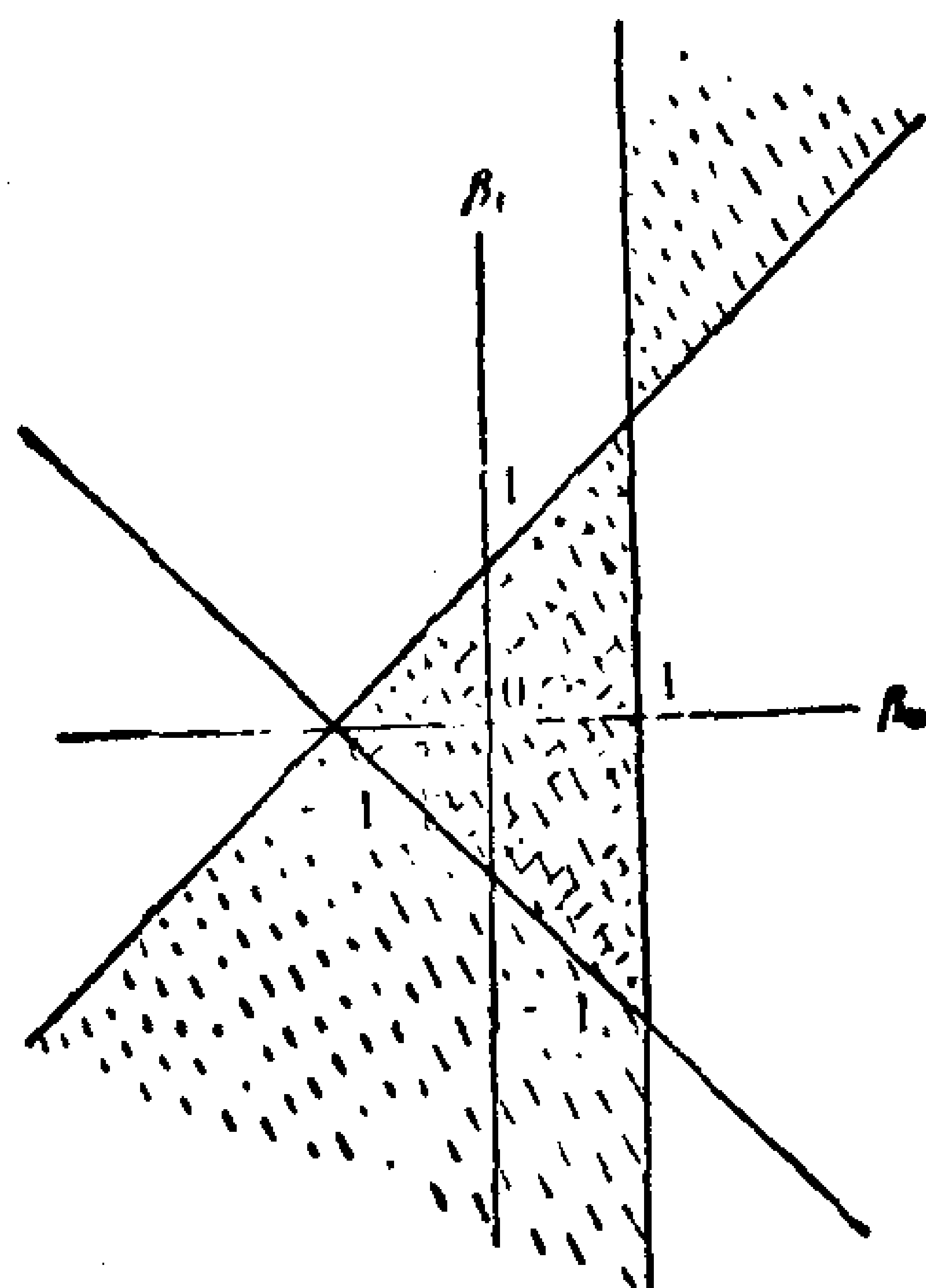


图 2.1　公式(2.20)为 A 稳定的三角形区域

因此条件(2.15)等价于 $\gamma(\beta_0, \beta_1) \geqslant 0$．方程

$$\gamma(\beta_0, \beta_1) \equiv (1-\beta_0)(1+\beta_0-\beta_1) = 0$$

表示退化的双曲线．满足条件（2.15）的点（β_0, β_1）的集合是图 2.1 中二个带阴影的楔形．满足（2.21）的开三角形是其中一个楔

形的真子集．因此条件(2.13)推出条件 (2.15)． 如果 (β_0, β_1)满足不等式(2.21)，则对应的公式(2.20)是 A 稳定的．

公式(2.20)的一个特例是向后差分公式

$$3y_{n+2} - 4y_{n+1} + y_n = 2hf_{n+2},$$

它对应于 $\beta_0 = \beta_1 = 0$．由上面的讨论，它是 A 稳定的．

下面来叙述 Dahlquist 对 A 稳定性的一个重要定理，它给出对 A 稳定线性多步公式的阶的一个限制性结果．

在(2.6)中取 $\phi(t) = e^{\lambda t}$，并令 $e^{\lambda h} = \zeta$，我们得到

$$\rho(\zeta) - \sigma(\zeta)\log\zeta = c_{p+1}(\log\zeta)^{p+1} + O((\lambda h)^{p+2}).$$

用 $\sigma(\zeta)$ 除等式两边，并作变换(2.17)—(2.19)，有

$$\frac{r(z)}{s(z)} = \log\frac{z+1}{z-1} + \frac{c_{p+1}}{\sigma(\zeta)}\left(\log\frac{z+1}{z-1}\right)^{p+1} + \frac{1}{\sigma(\zeta)}O((\lambda h)^{p+2}). \tag{2.22}$$

令

$$Z(z) = \frac{r(z)}{s(z)},$$

由定理 1 推出，若线性多步公式是 A 稳定的，则对任何具有 $\mathrm{Re}z > 0$ 的复数 z，将有 $\mathrm{Re}Z(z) > 0$．按照 Ozaki 和 Kasami 称具有这样性质的有理函数为正实有理函数．由 (2.22)，最高阶的 A 稳定的公式(2.2)可以用 $\log\dfrac{z+1}{z-1}$ 在 $z=\infty$ 处的最高阶的正实有理函数的近似给出．在无穷远点分别将 $\log\dfrac{z+1}{z-1}$ 和 $Z(z)$ 展开成 $\dfrac{1}{z}$ 的幂级数．得

$$\log\frac{z+1}{z-1} = \sum_{i=1}^{\infty} d_i\left(\frac{1}{z}\right)^i \tag{2.23}$$

和

$$Z(z) = e_{-m}z^m + e_{-(m-1)}z^{m-1} + \cdots + e_{-1}z + e_0 + \sum_{i=1}^{\infty} e_i\left(\frac{1}{z}\right)^i, \tag{2.24}$$

其中

$$d_i = \begin{cases} \dfrac{2}{i}, & i\text{为正奇数}, \\ 0, & \text{其它}. \end{cases}$$

而 e_i 是某个常数，且 $m < k$. 将这二个展开式代入(2.22)中，并略去 $\left(\frac{1}{z}\right)^{p+2}$ 阶的量，以及当 $z \to \infty$ 时 $\zeta \to 1$, $c^* = -c_{p+1}/\sigma(1)$，故得到

$$e_{-m}z^m + \cdots + e_{-1}z + e_0 + \sum_{i=1}^{p+1} e_i \left(\frac{1}{z}\right)^i$$

$$= \sum_{i=1}^{p+1} d_i \left(\frac{1}{z}\right)^i - c^* \left(\frac{2}{z}\right)^{p+1}.$$

比较等式二边的系数，我们得到

$$\begin{aligned} &e_i = 0, \quad \text{对于}\ i \leqslant 0, \\ &e_i = d_i, \quad \text{对于}\ 0 < i < p+1, \\ &e_{p+1} = d_{p+1} - 2^{p+1}c^*. \end{aligned} \tag{2.25}$$

由相容性条件，得

$$a_k = 0, \quad a_{k-1} = 2b_k.$$

不失一般性，我们假定 $b_k = 1$. 由(2.18)和(2.19)，$Z(z)$ 的定义及展开式(2.24)和系数(2.25)，得到 $r(z)$ 和 $s(z)$ 的系数之间有关系式

$$\begin{pmatrix} a_{k-1} \\ a_{k-2} \\ a_{k-3} \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} e_1 & & & & \\ e_2 & e_1 & & 0 & \\ e_3 & e_2 & e_1 & & \\ & & \ddots & \ddots & \\ e_k & e_{k-1} & e_{k-2} & \cdots & \end{pmatrix} \begin{pmatrix} 1 \\ b_{k-1} \\ b_{k-2} \\ \vdots \\ b_1 \end{pmatrix}. \tag{2.26}$$

因此，若公式的阶 $p \geqslant 2$, 则我们有

$$a_{k-1} = 2,$$

$$a_{k-2} = 2b_{k-1},$$

$$a_{k-3} = e_3 + 2b_{k-2}.$$

于是有理函数 $Z(z)$ 有形式

$$Z(z) = \frac{2z^{k-1} + 2b_{k-1}z^{k-2} + (e_3 + 2b_{k-2})z^{k-3} + \cdots}{z^k + b_{k-1}z^{k-1} + b_{k-2}z^{k-2} + \cdots}.$$

将其写成连分式，有

$$Z(z) = \cfrac{1}{\cfrac{z}{2} + \cfrac{1}{-\cfrac{4}{e_3}z + P_r}},$$

其中 P_r 是分子分母各为 $k-2$ 次的有理分式，在无穷远点处无极点. 由于 $Z(z)$ 是正实有理函数，量 $-\frac{4}{e_3}$ 不能是负的，e_3 不能等于 d_3，于是由(2.25)，公式的阶不能超过2. 若阶为 2，再由(2.25)得

$$e_3 = d_3 - 2^3c^* < 0,$$

因而得

$$c^* \geqslant \frac{1}{12}.$$

这就证明了 Dahlquist 的定理

定理 2.4 A 稳定的线性多步公式的阶不得超过 2，并且二阶的 A 稳定线性多步公式的误差常数以 $\frac{1}{12}$ 为下界.

令 $c^* \to \frac{1}{12}$，则 $Z(z)$ 退化成 $\frac{2}{z}$. 由于 $r(z)$ 和 $s(z)$ 无公因子，这些多项式除一个平凡的常数因子外由它们的商唯一确定. 对于梯形法

$$r(z)/s(z) = \rho(\zeta)/\sigma(\zeta) = 2(\zeta - 1)/(\zeta + 1) = 2/z,$$

其误差常数为 $c^* = \frac{1}{12}$，这就得到下面的推论.

推论 2.3 在所有的线性多步公式中，梯形法是唯一的阶为2，$c^* = \frac{1}{12}$ 的 A 稳定的公式.

§3 线性多步公式的 $A(\alpha)$ 稳定性

对于 $A(\alpha)$ 稳定性，Widlund[113] 证明了下述定理。

定理 2.5 (i) 显式的线性多步公式不能是 $A(0)$ 稳定的。

(ii) 存在唯一的 $p \geqslant k+1$ 的 $A(0)$ 稳定的方法，即梯形法，有

$$\rho(\zeta)=\zeta-1,\ \sigma(\zeta)=\frac{1}{2}(\zeta+1),\ p=2,$$

$$k=1,\ c^*=\frac{1}{12}.$$

定理 2.6 对于所有 $\alpha \in \left[0,\ \frac{\pi}{2}\right)$ 和 $k \leqslant 4$，存在 $A(\alpha)$ 稳定的 k 步 k 阶的线性多步公式。

由于 $A\left(\frac{\pi}{2}\right)$ 稳定的方法是 A 稳定的，当 $\alpha \to \frac{\pi}{2}$ 时，阶 $p \geqslant 3$ 的 $A(\alpha)$ 稳定公式的误差常数无限增大。

为了证明这二个定理，先证明下面二个引理。

引理 2.1 下面的命题是等价的

(i) 公式(2.2)是 $A(\alpha)$ 稳定的。

(ii) 对于所有 $h\lambda \in S_\alpha=\{z \mid |\arg(-z)|<\alpha, z \neq 0\}$ 方程

$$r(z)-h\lambda s(z)=0 \tag{2.27}$$

的根均在开左半复平面中。

(iii) 对于 $\operatorname{Re} z>0$，$r(z)/s(z)$ 是正则的，并且在 S_α 的补中取值。

证明 (i) 和 (ii) 的等价性直接由 $A(\alpha)$ 稳定性的定义给出。命题 (ii) 说明方法(2.2)是 $A(\alpha)$ 稳定等价于当(2.27)的根 z 有 $\operatorname{Re} z \geqslant 0$ 时，$h\lambda$ 在 S_α 的补中。为了完成由 (ii) 推出 (iii)，仅须证明，如果 $\operatorname{Re} z>0$，则 $s(z)$ 不能为零。假定 $\operatorname{Re} z_1>0$，但有 $s(z_1)=0$，$r(z_1) \neq 0$ (因为 $r(z)$ 和 $s(z)$ 无公因子)，于是对某个

$c_1 \neq 0$ 和 $m > 0$ 成立

$$r(z)/s(z) \sim c_1(z - z_1)^{-m} \quad (z \to z_1),$$

这表示 $r(z)/s(z)$ 将对某个 $\operatorname{Re} z > 0$ 的 z 取 S_α 中的值，这与 (ii) 矛盾.

由于 S_α 是开集，容易由 (iii) 推出 (ii)，引理证毕.

由(2.22)，得

$$\log \frac{z+1}{z-1} - \frac{r(z)}{s(z)} \sim c^* \left(\frac{2}{z}\right)^{p+1} (z \to \infty), \qquad (2.28)$$

再由展开式(2.23)，并且比较

$$(a_0 + a_1 z + \cdots + a_k z^k) - 2(b_0 + b_1 z + \cdots + b_k z^k)$$
$$\times \left[\frac{1}{z} + \frac{1}{3}\left(\frac{1}{z}\right)^3 + \cdots + \frac{1}{2\mu+1}\left(\frac{1}{z}\right)^{2\mu+1} + \cdots\right]$$
$$\sim c_{p+1}\left(\frac{2}{z}\right)^{p-k+1} \quad (z \to \infty).$$

令同次幂的系数和相等，并令 $b_j = 0,\ j > k$，可得对于

$$\max(0, k-p) \leqslant \nu \leqslant k,$$

有

$$a_\nu = \sum_{\mu \geqslant 0} 2b_{\nu+1+2\mu}/(2\mu+1). \qquad (2.29)$$

当 $p < k$ 时

$$a_{k-p-1} = \sum_{\mu \geqslant 0} 2b_{k-p+2\mu}/(2\mu+1) - c^* b_k 2^{p+1}, \qquad (2.30)$$

而当 $p \geqslant k$ 时，有

$$0 = \sum_{\mu \geqslant 0} 2b_{\nu+2\mu}/(2\mu+1), \quad 0 \leqslant \nu \leqslant p-k-1, \quad (2.31)$$

$$0 = \sum_{\mu \geqslant 0} 2b_{p-k+2\mu}/(2\mu+1) - c^* b_k 2^{p+1}. \qquad (2.32)$$

引理 2.2 若公式(2.2)是 $A(0)$ 稳定的，则有

$$a_\nu \geqslant 0,\ b_\nu \geqslant 0,\ \nu = 0, 1, \cdots, k.$$

证明　对于实系数的多项式，若它的根均在左半平面内部，则

其系数一定有相同的符号．对(2.27)分别取$h\lambda=0$和$h\lambda=-\infty$，应用引理 2.1 的 (ii)，立即可得本引理的结论．

定理 2.5 的证明　如果 $\beta_k=0$，则由(2.19)推出

$$\sum_{\nu=0}^{k} b_\nu=0.$$

由引理 2.2，若公式是 $A(0)$ 稳定的，$b_k=0$．于是由于 $\sigma(1)=2^k b_k$，得 $\sigma(1)=0$，这与 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 无公因子相矛盾，所以定理的 (i) 成立．

如果 $p>k+1$，则由(2.31)和引理 2.2 仍可推出 $b_k=0$．也得到与上述相同的矛盾．如果 $p=k+1$，在(2.31)中取 $\nu=0$，由引理 2.2 推得

$$b_0=b_2=b_4=\cdots=0,$$

所以 $s(z)$ 是奇函数．但是由(2.29)，$r(z)$ 是偶函数．由 Henrici[62] 推出，当 $k>1$ 时是弱稳定的，$\rho(\zeta)$ 的根均在单位圆上，并且(2.10)的根可表成

$$\zeta_{jq}=\zeta_j(1+k_j q+O(|q|^2)),$$

其中 ζ_j 是 $\rho(\zeta)=0$ 的根，k_j 是增长参数，$q=h\lambda$．这时有 $k_j<0$，因此当 q 是负数，并且 $|q|$ 很小时，将会有 $|\zeta_{jq}|>1$，这与 $A(0)$ 稳定性相矛盾．唯一的例外情形为 $k=1$，我们得到

$$r(z)=2b_1,\ s(z)=b_1 z,$$

这是与梯形法相应的多项式，于是定理的 (ii) 证毕．

定理 2.6 的证明　由引理 2.1 的 (iii)，通过直接计算可以证明下面的结论成立．

对于 $k=p=3$，如果

$$b_0>0,\ b_2>3^{-\frac{1}{2}},\ b_3=1,$$

$$b_1>\max\left(b_0/b_2,\ 3b_0b_2,\ (b_0+b_2/3)^3\mathrm{tg}^2\alpha\left(b_0\left(b_2^2-\frac{1}{3}\right)\right)\right),$$

则方法是 $A(\alpha)$ 稳定的．

对于 $k=p=4$，如果取

$$b_4=1,\ b_3>3^{-\frac{1}{2}},\ b_0>0,$$

$$b_1 = b_3(b_2/2 + (b_2^2/4 - b_0)^{\frac{1}{2}} - 1),$$

并且 b_2 取得充分大，则方法对于指定的 $\alpha \in \left(0, \frac{\pi}{2}\right)$ 是 $A(\alpha)$ 稳定的.

下面将给出判定公式(2.2)为 $A(\alpha)$ 稳定的一个充分条件，这个条件是定理 2.3 对 $A(\alpha)$ 稳定性的推广．先给出几个引理．令 $q = h\lambda$, $\zeta_i(q)$, $i = 1, \cdots, k$ 是(2.10)的根．由定义 1.5 立即可得下面的引理.

引理 2.3 公式(2.2)是 $A(\alpha)$ 稳定的充要条件为当 $q \in S_\alpha$ 时有

$$|\zeta_i(q)| < 1, \quad i = 1, \cdots, k.$$

引理 2.4 如果公式(2.2)是 $A(\alpha)$ 稳定的，则有

$$|\sigma_i| \leqslant 1, \quad i = 1, 2, \cdots, k,$$

其中 σ_i 是多项式 $\sigma(\zeta)$ 的零点.

引理 2.5 假设成立条件

(i) $|\sigma_i| < 1, i = 1, 2, \cdots, k$

和

(ii) 当 $|\zeta| \geqslant 1$ 时，有 $q(\zeta) = \rho(\zeta)/\sigma(\zeta) \notin S_\alpha$.

则公式(2.2)是 $A(\alpha)$ 稳定的.

由定理 2.5 的 (i)，只有隐式方法(即 $\beta_k \neq 0$) 才可能是 $A(\alpha)$ 稳定的．因此，我们仅讨论隐式的情形.

定理 2.7 设下列条件成立

(i) 公式(2.2)是零稳定的，即 $\rho(\zeta)$ 的零点的模均不超过 1，并且模为 1 的零点均是单的，

(ii) $\beta_k \neq 0$,

(iii) $\operatorname{Im} q(e^{i\theta}) \geqslant 0, \theta \in [0, \pi]$,

(iv) $\operatorname{Im} q(e^{i\theta}) + \operatorname{tg} \alpha \operatorname{Re} q(e^{i\theta}) \geqslant 0, \theta \in [0, \pi]$,

(v) $|\sigma_i| < 1, i = 1, \cdots, k$,

(vi) $\alpha_k/\beta_k > 0$.

则公式(2.2)是 $A(\alpha)$ 稳定的.

证明 设条件 (i) 到 (vi) 成立,则在 $|\zeta|\geqslant 1$ 上, $q(\zeta)$ 是解析的,而 $\mathrm{Re}\, q(\zeta)$ 和 $\mathrm{Im} q(\zeta)$ 是调和的.

为了证明公式(2.2)是 $A(\alpha)$ 稳定的,我们应用引理 2.5. 由于若 $q(\zeta)\notin S_\alpha$, 则有 $q(\bar\zeta)\notin S_\alpha$, 其中 $\bar\zeta$ 是 ζ 的共轭复数, 以及若 $q(\zeta)\notin S_\alpha$, 则有

$$|\mathrm{Im} q(\zeta)| + \mathrm{tg}\,\alpha \mathrm{Re} q(\zeta) \geqslant 0. \tag{2.33}$$

为了完成定理的证明,只须证明当 $\zeta\in D=\{\zeta\,|\,|\zeta|\geqslant 1 \text{ 和 } \mathrm{Im}\zeta\geqslant 0\}$ 时(2.33)成立. 因为这就保证了 $q(\zeta)=\rho(\zeta)/\sigma(\zeta)\notin S_\alpha$. 由于 $\mathrm{Im} q(\zeta)$ 在 D 中是调和的,以及当 ζ 为实数时, $\mathrm{Im} q(\zeta)=0$, 由条件(iii)和调和函数的极小值原理推出当 $\zeta\in D$ 时, $\mathrm{Im} q(\zeta)\geqslant 0$.

由于函数

$$u(\zeta)=\mathrm{Im} q(\zeta)+\mathrm{tg}\,\alpha \mathrm{Re} q(\zeta)$$

是 D 中的调和函数,再对 $u(\zeta)$ 应用极值原理,我们得到

$$u(\zeta)\geqslant \min_{\zeta\in\partial D} u(\zeta),$$

其中 ∂D 是 D 的边界

$$\partial D=\{\zeta\,|\,\zeta=e^{i\theta},\ \theta\in[0,\pi]\}\cup(-\infty,-1]\cup[1,\infty).$$

由于条件 (i), (iv), (v) 和 (vi) 成立,当 $\zeta\in D$ 时,有 $u(\zeta)\geqslant 0$. 定理证毕.

条件 (i), (ii), (v), (vi) 是容易验证的,而 (iii), (iv) 稍困难一点. 如果 (v) 成立,则 $\sigma(e^{i\theta})\neq 0$. $\theta\in[0,\pi]$. 因此 (iii) 和 (iv) 又分别等价于

$$\mathrm{Im}\,\tilde q(e^{i\theta})\geqslant 0,\ \theta\in[0,\pi], \tag{2.34}$$

$$\mathrm{Im}\,\tilde q(e^{i\theta})+\mathrm{tg}\,\alpha \mathrm{Re}\,\tilde q(e^{i\theta})\geqslant 0.\ \theta\in[0,\pi], \tag{2.35}$$

其中

$$\tilde q(e^{i\theta})=\rho(e^{i\theta})\sigma(e^{-i\theta})$$

$$=\sum_{j=0}^{k}\gamma_j\cos(j\theta)+i\sum_{j=1}^{k}\delta_j\sin(j\theta),$$

$$\gamma_0=\sum_{l=0}^{k}\alpha_l\beta_l,$$

$$\gamma_j = \sum_{l=0}^{k-j} (\alpha_{l+j}\beta_l + \alpha_l\beta_{l+j}),\ j = 1, 2, \cdots, k,$$

$$\delta_j = \sum_{l=0}^{k-j} (\alpha_{l+j}\beta_l - \alpha_l\beta_{l+j}),\ j = 1, 2, \cdots, k.$$

这样，如果条件 (v) 成立，则条件 (iii)(iv) 分别等价于

$$I_k(x) \geqslant 0,\ x \in [-1, 1], \tag{2.36}$$

$$\sqrt{1 - x^2} \cdot I_k(x) + \mathrm{tg}\,\alpha R_k(x) \geqslant 0,\ x \in [-1, 1],$$

其中

$$I_k(x) = \sum_{j=1}^{k} \delta_j U_j(x),$$

$$R_k(x) = \sum_{j=0}^{k} \gamma_j T_j(x)$$

而

$$T_j(x) = \cos j\theta,\ U_j(x) = \frac{\sin j\theta}{\sin\theta},\ x = \cos\theta$$

为 Чебышев 多项式.

例 2.3　对下面的公式应用定理 2.7，寻找最大的 $\alpha \in \left[0, \frac{\pi}{2}\right)$，使公式是 $A(\alpha)$ 稳定的.

$k = 3,\ 11y_n - 18y_{n-1} + 9y_{n-2} - 2y_{n-3} = 6hf_n,$

$k = 4,\ 25y_n - 48y_{n-1} + 36y_{n-2} - 16y_{n-3} + 3y_{n-4} = 12hf_n,$

$k = 5,\ 137y_n - 300y_{n-1} + 300y_{n-2} - 200y_{n-3} + 75y_{n-4} - 12y_{n-5} = 60hf_n,$

$k = 6,\ 147y_n - 360y_{n-1} + 450y_{n-2} - 400y_{n-3} + 225y_{n-4} - 72y_{n-5} + 10y_{n-6} = 60hf_n.$

对于这些公式条件 (i), (ii), (v) 和 (vi) 均满足. 相应的 $R_k(x)$ 和 $I_k(x)$ 分别为

$R_3(x) = -6(x - 1)^2(4x - 1),$

$R_4(x) = 96(x - 1)^3(3x + 1),$

$$R_5(x) = -480(x-1)^3(24x^2-3x-11),$$
$$R_6(x) = 480(x-1)^4(40x^2+16x-11),$$
$$I_3(x) = 12(4x^2-9x+8),$$
$$I_4(x) = -48(6x^3-16x^2+15x-18),$$
$$I_5(x) = 120(96x^4-300x^3+328x^2-150x+56),$$
$$I_6(x) = -240(80x^5-288x^4+370x^3-184x^2+15x-8).$$

通过计算 $I_k(x)(k=3,4,5,6)$ 的根，得到

$$I_k(x) \geqslant 0, \quad k=3,4,5,6, \quad x \in [-1, 1]$$

即条件（iii）成立．再由

$$\operatorname{tg}(\alpha_{\max}) = \min_{x \in A_k}\left(-\frac{\sqrt{1-x^2}\cdot I_k(x)}{R_k(x)}\right), \quad k=3,4,5,6.$$

确定使（iv）成立的最大的 $\alpha \in \left[0, \frac{\pi}{2}\right)$，其中

$$A_k = \{x \mid x \in [-1, 1],\ R_k(x) < 0\}.$$

下面的表是得到的 $\alpha_{\max}$ 的值（误差为 1′）

表 2.1

k	3	4	5	6
$\alpha_{\max}$	88°27′	73°14′	51°50′	18°47′

§4 线性多步公式的 A_0 稳定性

令

$$f(z,q) = r(z) - qs(z),$$
$$H_+ = \{z \mid \operatorname{Re} z > 0\},$$
$$H_- = \{z \mid \operatorname{Re} z < 0\},$$
$$D = \{\zeta \mid |\zeta| < 1\}.$$

任意集合 S 的闭包和边界分别用 $\bar{S}$ 和 ∂S 表示．特别 ∂H_- 是虚轴．复数 z 的共轭复数记成 $\bar{z}$．

下面的定理用函数 $f(z,q)$ 来刻划 A_0 稳定公式．

定理 2.8 公式(2.2)是 A_0 稳定的等价于对所有 $q\in(-\infty, 0)$，$f(z, q)$ 的零点均在 H_- 中.

定理 2.9 下面的命题是等价的.

(i) 公式(2.2)是 A_0 稳定的.

(ii) 对于 $z\in H_+$，$r(z)/s(z)$是正则的，并且不取$(-\infty, 0)$中的值. 对于 $z\in\partial H_+$，$r(z)\overline{s(z)}$不取$(-\infty, 0)$中的值.

(iii) 对于$\overline{D}$的补中的 ζ，$\rho(\zeta)/\sigma(\zeta)$ 是正则的，不取$(-\infty, 0)$中的值. 对于 $\zeta\in\partial D$，$\rho(\zeta)\overline{\sigma(\zeta)}$不取$(-\infty, 0)$中的值.

定理 2.10 若公式(2.2)是 A_0 稳定的，则多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 的零点在 $\overline{D}$ 中，多项式 $r(z)$和 $s(z)$ 的零点在 $\overline{H}_-$ 中；另外 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 在 ∂D 上的任何零点和 $r(z)$和 $s(z)$在 ∂H_- 上的任何零点最多是二重的.

证明 只证明 $\sigma(\zeta)$ 在 ∂D 上的任何零点最多是二重的. 设 ζ_1 是 $\sigma(\zeta)$ 在 ∂D 上的一个零点，则有

$$q(\zeta)=\frac{\rho(\zeta)}{\sigma(\zeta)}\sim c(\zeta-\zeta_1)^{-m}\ (\zeta\to\zeta_1),$$

其中 $c\neq 0$. 若 $m\geqslant 3$. 则当 ζ 在D的外部的以 ζ_1 为心的充分小的圆内连续变化时，$q(\zeta)$ 的幅角将连续地至少变化 $3\pi-\alpha$. 因而当 α 充分小时，$q(\zeta)$ 将会取到 $(-\infty, 0)$中的值. 这与定理 2.9 的 (iii) 矛盾，所以 $m\leqslant 2$.

下面的定理给出对于 A_0 稳定公式的系数的限制.

定理 2.11 如果公式(2.2)是 A_0 稳定的，则对于 $0\leqslant j\leqslant k$，有

$$a_j\geqslant 0,\ b_j\geqslant 0.$$

另外，有

$$\sum_{j=0}^{k}a_j>0,\ \sum_{j=0}^{k}b_j>0,\ \beta_k>0,\ b_k>0.$$

由这定理推出，A_0 稳定的线性多步公式一定是隐式的.

如果公式(2.2)的阶 $p\geqslant 1$. 则当 $k-p\leqslant\nu\leqslant k$ 时，a_ν 由(2.29)给出. 我们按以前的约定，当 $\nu<0$ 时，$a_\nu=0$，和当 $\nu>k$ 时，$b_\nu=0$.

定理 2.12 设公式(2.2)是 A_0 稳定的，并且 $k \geqslant 3$, $p \geqslant 3$, 则对于 $2 \leqslant j \leqslant k$ 有 $b_j > 0$; 对于 $\max(0, k-p) \leqslant j \leqslant k-1$ 有 $a_j > 0$; 对于 $2 \leqslant j \leqslant k-1$, 有 $b_j > 0$.

证明 假定 $b_{k-1} = 0$. 由于当 $q \in (-\infty, 0)$ 时，$f(z, q)$ 的零点均在 H_- 中，应用 Routh-Hurwitz 准则，对应的 Hurwitz 行列式 Δ_k 必须是正的. 应用(2.29)得

$$a_k = 0, a_{k-1} = 2b_k,\ a_{k-2} = 2b_{k-1},\ a_{k-3} = 2b_{k-2} + \frac{2}{3} b_k,$$

所以

$$\Delta_2 = \begin{vmatrix} a_{k-1} - qb_{k-1} & a_k - qb_k \\ a_{k-3} - qb_{k-3} & a_{k-2} - qb_{k-2} \end{vmatrix}$$

$$= \begin{vmatrix} 2b_k & -qb_k \\ 2b_{k-2} - qb_{k-3} + \dfrac{2}{3} b_k & -qb_{k-2} \end{vmatrix}$$

$$= qb_k\left(\frac{2}{3} b_k - qb_{k-3}\right) < 0,$$

得到矛盾. 因而 $b_{k-1} > 0$.

由于 $s(z)$ 的零点在 $\bar{H}_-$ 中，因此可将其表成

$$s(z) = b_k z^m s_1(z) s_2(z),$$

其中 $s_1(z)$ 含有 $s(z)$ 的所有虚的零点，$s_2(z)$ 含有 $s(z)$ 的所有具有严格负实部的零点. 容易看出，$s_1(z)$ 的偶次项系数是非零的，而 $s_2(z)$ 的所有系数是非零的，并具有相同的符号. 由于 $b_k \neq 0$ 和 $b_{k-1} \neq 0$, 推出 $s_2(z)$ 的次数至少为 1. 得出 $s_1(z)s_2(z)$ 的所有系数是非零的. 另外，由定理 2.10, $s(z)$ 在 $z = 0$ 处最多有二重零点，所以 $m \leqslant 2$. 因此对于 $2 \leqslant j \leqslant k$, 有 $b_j > 0$.

由 (2.29), 立即得到对于 $\max\{0, k-p\} \leqslant j \leqslant k-1$, 有 $a_j > 0$. 由于 $a_{k-1} > 0$ 和 $a_{k-2} > 0$, 应用类似于对系数 b_j 的证明，可得到对于 $2 \leqslant j \leqslant k-1$, 有 $a_j > 0$. 定理证毕.

通过与 § 2 的定理 2.5 的类似证明，可得

定理 2.13 只存在唯一的阶 $p \geqslant k+1$ 的 A_0 稳定的线性多

步公式，即梯形法.

例 2.4 如果多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 对应的公式(2.2)是收敛的，则 $\rho(\zeta)$ 在 ∂D 上只有单根. 考虑由

$$\rho(\zeta)=\zeta^2-\zeta,\sigma(\zeta)=(\zeta+1)^2/4 \tag{2.37}$$

所对应的公式(2.2). 它是收敛的. 方程(2.10)的零点为

$$\zeta=\frac{4+2h\lambda\pm4\sqrt{1+2h\lambda}}{8-2h\lambda}.$$

易证当 $\lambda\in(-\infty,0)$ 时，$|\zeta|<1$. 因此它是 A_0 稳定的. 这个例子说明定理 2.10 已不能再加强成要求多项式 $\sigma(\zeta)$ 在 ∂D 上的零点是一重的.

另外，通过直接计算，可知 (2.37) 所对应的方法 (2.2) 不是 $A(0)$ 稳定的. 这样可得出 A_0 稳定的方法类包含 $A(0)$ 稳定的方法类，并且两者是不重合的.

例 2.5 如果 $d\geqslant 2^{k+1}$，则对应于多项式

$$s(z)=(z+d)^k \tag{2.38}$$

的线性 k 阶 k 步公式是 A_0 稳定的，多项式 $r(z)$ 的系数由(2.29)确定. 这个方法称作 Cryer 方法.

证明 由于 $s(z)$ 的零点在 H_- 中和 $f(z,q)$ 的零点是 q 的连续函数，应用定理 2.8，只要证明对于 $q\in(-\infty,0)$，$f(z,q)$ 无虚的零点，便可知道当 $q\in(-\infty,0)$ 时，$f(z,q)$ 的零点都在 H_- 中.

通过直接计算，可以得到

$$f(dw;q)=d^{k-1}[T_0(w)+T_1(w)],$$

其中

$$T_0(w)=[(-qdw+2)(1+w)^k-2]/w,$$

$$T_1(w)=2\sum_{m=0}^{k-1}w^m\sum_{i>0}\binom{k}{m+1+2j}d^{-2j}/(1+2j).$$

这里约定如果 $l>k$ 时，$\binom{k}{l}=0$.

进行初等的估计，可以得到估计式

$$|T_0(iy)| \geqslant \begin{cases} 2y^{k-1}, & y \in [1, \infty), \\ \dfrac{1}{4}, & y \in \left[\dfrac{1}{2^k}, 1\right], \\ \dfrac{k}{2}, & y \in \left(0, \dfrac{1}{2^k}\right] \end{cases}$$

和

$$|T_1(iy)| \leqslant \frac{1}{12} \max[1, y^{k-1}], \ y \in [0, \infty),$$

比较这二个估计式，对所有 $y \in [0, \infty)$，有

$$|T_0(iy)| > |T_1(iy)|.$$

所以，$f(z,q)$ 无虚的零点．证完．

由这个例子知道任意高阶的 A_0 稳定的线性多步公式是存在的．

例 2.6 $k \geqslant 2$ 的 k 步 Adams-Moulton 方法不是 A_0 稳定的．

对于这些隐式方法，由 $\sigma(\zeta)$的系数的显式表示式(见 Gear[20]) 可知，多项式 $r(z)$ 的系数 b_0 和 b_k 有不同的符号． 所以由定理 2.11 推得 $k \geqslant 2$ 的 Adams-Moulton 方法不具有 A_0 稳定性．

下面利用 A_0 稳定性及多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 的一些根条件来给出 $A(0)$ 稳定性的一个充分必要条件．

定理 2.14 收敛的公式(2.2)是 $A(0)$ 稳定的充分必要条件是

(i) 公式(2.2)是 A_0 稳定的，

(ii) 多项式 $\sigma(\zeta)$ 的模为 1 的根是单根，

(iii) 若 ζ 为 $\rho(\zeta)$ 的根，且$|\zeta| = 1$，有

$$\mathrm{Re}\left(\frac{\sigma(\zeta)}{\zeta\rho'(\zeta)}\right) > 0,$$

(iv) 若 ζ 为 $\sigma(\zeta)$ 的根，且$|\zeta| = 1$，则有

$$\mathrm{Re}\left(\frac{\rho(\zeta)}{\zeta\sigma'(\zeta)}\right) > 0.$$

证明 必要性 (i) 由 $A(0)$ 稳定性和 A_0 稳定性的定义立

即推得．令ξ是多项式$\sigma(\zeta)$的模为1的根． 由于ρ和σ无公因子，在ξ的邻域中有

$$q(\zeta)=\rho(\zeta)/\sigma(\zeta)\sim c(\zeta-\xi)^{-m}, \qquad (2.39)$$

其中m是某个正整数．由于公式(2.2)是$A(0)$稳定的，所以存在$\alpha\in\left(0,\ \frac{\pi}{2}\right)$，使当$|\zeta|>1$时，有$|\arg(-q(\zeta))|\geqslant\alpha$． 如果$m\geqslant2$，我们可以选取以$\xi$为心的圆，使得如果沿这个圆，且在单位圆外的开部分连续移动时，$q(\zeta)$的幅角将至少连续地变化$2\pi-\alpha$．这对于选取的α，公式为$A(\alpha)$稳定的矛盾，所以(ii)成立．

令$\zeta(q)$是(2.10)的根，$q=h\lambda$，且有$|\zeta(0)|=1$、由待定系数法，我们找到

$$\zeta(q)=\xi\left[1+\frac{\sigma(\xi)}{\xi\rho'(\xi)}q+O(q^2)\right], \qquad (2.40)$$

其中$\xi=\zeta(0)$． 由于$\sigma(\xi)\neq0$，可以记$\sigma(\xi)/(\xi\rho'(\xi))=de^{i\varphi}$，$d>0$．显然，如果$\varphi\in\left(\frac{\pi}{2},\ \frac{3}{2}\pi\right)$，方法不是$A_0$稳定的，因而也不是$A(0)$稳定的．令$\varphi=\frac{3\pi}{2}$． 由方法是$A(0)$稳定的，存在$\alpha>0$，使对所有$\mu>0$有$|\zeta(\mu e^{i(\pi-\frac{\alpha}{2})})|<1$．但由(2.40)，对充分小的$\mu$得到

$$|\zeta(\mu e^{i(\pi-\frac{\alpha}{2})})|=|1+\mu de^{i(\frac{\pi}{2}-\frac{\alpha}{2})}+O(\mu^2)|>1.$$

因此$\varphi\neq\frac{3}{2}\pi$．类似地，可得到$\varphi\neq\frac{\pi}{2}$．所以(iii)成立．

现在设$|\zeta(\infty)|=1$，$\zeta(\infty)$是多项式$\sigma(\zeta)$的零点．由待定系数法，我们得到

$$\zeta(q)=\hat{\xi}\left[1+\frac{\rho(\hat{\xi})}{\hat{\xi}\sigma'(\hat{\xi})}\cdot\frac{1}{q}+O\left(\frac{1}{q^2}\right)\right],$$

其中$\hat{\xi}=\zeta(\infty)$．与证明(iii)的必要性类似，可得(iv)的必要性．

充分性　设收敛的线性多步公式是 A_0 稳定的，它的多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 满足条件 (ii)—(iv)．由方程(2.10)定义的 $\zeta(q)$ 是 q 的 k 值函数．为证明公式是 $A(0)$ 稳定的，必须证明存在一个 $\alpha\in\left(0,\ \frac{\pi}{2}\right)$，使只要 $q\in S_\alpha$，$\zeta(q)$ 的所有 k 个值均将满足 $|\zeta(q)|<1$．由于 ρ 和 σ 的系数是实数，有 $\zeta(\bar{q})=\overline{\zeta(q)}$，再由条件 (i)，所以我们可只限于讨论 $\mathrm{Im}q>0$ 的情形．

$\zeta(q)$ 是一个代数函数，仅有有限多个支点和极点．因此，存在正常数 α^*，r^* 和 R^*，使在区域 $\Phi=D_0^*\cup S_{\alpha^*}^+\cup D_\infty^*$ 中 $\zeta(q)$ 无支点和极点，其中

$$S_{\alpha^*}^+=\{q\,|-\alpha^*<\arg(-q)<0,\ q\neq 0\},$$
$$D_0^*=\{q\,||q|<r^*,\ \mathrm{Im}q>0\},$$
$$D_\infty^*=\{q\,||q|>R^*,\ \mathrm{Im}q>0\},$$

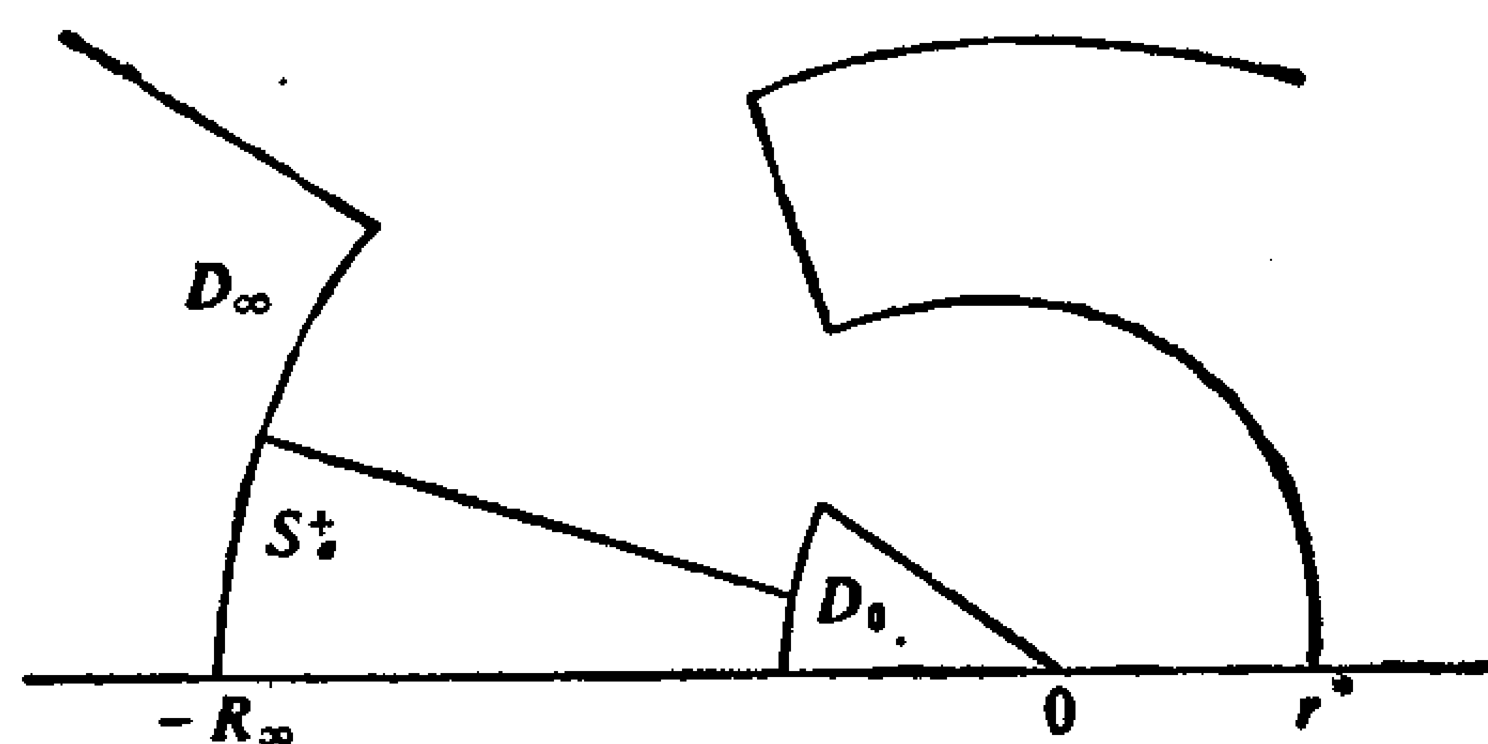


图　2.2

Φ 是一个单连通区域，$\zeta(q)$ 在 Φ 中可以表成 k 个分支，$\zeta_i(q)$，$j=1,2,\cdots,k$，其中每个 $\zeta_i(q)$ 在 Φ 中是解析的．显然，我们只需证明对每个分支 $\zeta_i(q)$，存在正数 α_i $(0<\alpha_i<\alpha^*)$，使只要 $q\in S_{\alpha_j}^+=\{q\,|-\alpha_i<\arg(-q)<0,\ q\neq 0\}$ 就有 $|\zeta_i(q)|<1$．

下面仅考虑一个分支，为方便起见，省掉下标．令

$$\lim_{q\to 0}|\zeta(q)|<1,$$

则存在一个常数 r_0，$0<r_0<r^*$，使对所有 $q\in\{q\in\Phi\,||q|\leqslant r_0\}$ 有 $|\zeta(q)|<1$．令 $\lim\limits_{q\to 0}|\zeta(q)|=1$，由于方法是收敛的，

$$\lim_{q\to 0}\zeta(q)=\xi$$

是 $\rho(\zeta)$ 的单根. $\zeta(q)$ 在以 $q=0$ 为心,半径充分小的圆盘上是解析的. 在这个圆盘上, $\zeta(q)$ 有表示式(2.40),即

$$\zeta(\mu e^{i\eta})=\xi[1+\mu d e^{i(\eta+\varphi)}+O(\mu^2)],$$

其中 $de^{i\varphi}=\sigma(\xi)/(\xi\rho'(\xi))$, $d>0$ 和 $q=\mu e^{i\eta}$, $\mu\geqslant 0$. 因此有

$$\begin{aligned}|\zeta(\mu e^{i\eta})|=|\xi||1+\mu d\cos(\eta+\varphi)+i\mu d\sin(\eta+\varphi)\\ +O(\mu^2)|=(1+2\mu d\cos(\eta+\varphi)\\ +O(\mu^2))^{\frac{1}{2}},\end{aligned}$$

由于 $\varphi\in\left(-\dfrac{\pi}{2},\dfrac{\pi}{2}\right)$, 存在 α_0, $0<\alpha_0<\alpha^*$, 使对所有具有 $|\pi-\eta|\leqslant\alpha_0$ 的 η, 有 $\cos(\eta+\varphi)<0$. 因此总存在正常数 $r_0<r^*$ 和 $\alpha_0<\alpha^*$, 使得只要 $q\in D_0=\{q\in\Phi\,|\,|q|\leqslant r_0, |\arg(-q)|<\alpha_0\}$, 就有 $|\zeta(q)|<1$. 对点 $q=\infty$ 进行类似的讨论,可以找到正常数 $R_\infty>R^*$ 和 α_∞, $0<\alpha_\infty<\alpha^*$, 使只要

$$q\in D_\infty=\{q\in\Phi\,|\,|q|\geqslant R_\infty, |\arg(-q)|<\alpha_\infty\},$$

就有 $|\zeta(q)|<1$. 由条件(i),存在 $\zeta(q)$ 在 $\Phi\cup[-R_\infty,-r_0]$上的连续延拓,并且只要 $q\in[-R_\infty,-r_0]$就有 $|\zeta(q)|<1$. 因此存在正常数 $\alpha_l<\alpha^*$, 使对所有

$$q\in S^+_{\alpha_l}=\{q\in\Phi\,|\,|\arg(-q)|<\alpha_l, |q|>r_0, |q|<R_\infty\},$$

有 $|\zeta(q)|<1$. 取 $\alpha=\min\{\alpha_0,\alpha_l,\alpha_\infty\}$, 于是对所有 $q\in S_\alpha$ 成立 $|\zeta(q)|<1$, 这表示方法是 $A(0)$ 稳定的. 定理证毕.

注 2.1 定理的条件 (i) 不能推出条件 (ii), 因为例 2.4 中的方法是 A_0 稳定的,但其中 $\sigma(\zeta)$ 具有模为 1 重数为 2 的根. 定理 2.10 只说明 A_0 稳定的方法的多项式 $\sigma(\zeta)$ 不能有模为 1 而重数超过 2 的零点.

注 2.2 由条件(i)推出,只要 ξ 是 $\rho(\zeta)$ 的零点,且有 $|\xi|=1$, 则

$$\operatorname{Re}\left(\sigma(\xi)/(\xi\rho'(\xi))\right)\geqslant 0. \tag{2.41}$$

由多项式

$$\rho(\zeta)=(\zeta-1)(\zeta^2+1),$$
$$\sigma(\zeta)=2\zeta^3+\zeta-1$$

所确定的方法是 A_0 稳定的，并且满足条件 (ii) 和 (iv)，但不是 $A(0)$ 稳定的．原因是对这个方法恰好出现(2.41)中的等号．所以条件 (i) 不能推出 (iii)．

注 2.3　与注 2 类似，只要 $\bar{\zeta}$ 是 $\sigma(\zeta)$ 的模为 1 的根，则由条件 (i) 只能推出

$$\operatorname{Re}(\rho(\bar{\zeta})/(\bar{\zeta}\sigma'(\bar{\zeta}))) \geqslant 0, \tag{2.42}$$

还不能推出条件 (iv)．考虑由多项式

$$\rho(\zeta) = \zeta^3 - \frac{4}{3}\zeta^2 + \zeta - \frac{2}{3},$$

$$\sigma(\zeta) = \frac{2}{3}\zeta(\zeta^2 + 1)$$

所确定的方法．它是 A_0 稳定的，满足条件 (ii)，但不是 $A(0)$ 稳定的．这是因为(2.42)中出现等号．

例 2.7　考虑例 2.5 中的多项式(2.38)．它对应的 $\sigma(\zeta)$ 为

$$\sigma(\zeta) = (1 + d)^k\left(\zeta + \frac{1 - d}{1 + d}\right),$$

当 $d \geqslant 2^{k+1}$ 时，$\dfrac{1 - d}{1 + d}$ 满足不等式

$$-1 < \frac{1 - d}{1 + d} \leqslant -1 + \frac{2}{(1 + 2^{k+1})},$$

这表示 $\sigma(\zeta)$ 只具有模小于 1 的零点．再由例 2.5 中的证明，立即可推得多项式 $\rho(\zeta)/(\zeta - 1)$ 只具有模小于 1 的零点．而对于 $\zeta = 1$，定理 2.14 的条件 (iii) 满足．由定理 2.14，这个方法是 $A(0)$ 稳定的．

在[46]中，Cryer 提出这样一个猜测，存在一个正整数 N，将不存在阶大于 N 的 $A(0)$ 稳定的方法．但 Cryer 方法是 $A(0)$ 稳定的 k 步 k 阶的方法，所以这个猜测不成立．由定理 2.13，当 $k \geqslant 2$ 时，$A(0)$ 稳定的方法的误差阶 p 有界 $p \leqslant k$．由这个例子看到，这个界是可以达到的．

§5 线性多步公式的刚性稳定性

下面的定义比定义 1.7 稍精确一点. 设 D, θ, a 是正常数,定义集合

$$R_1 = \{q \mid \mathrm{Re}\, q < -D\},$$
$$R_2 = \{q \mid \mathrm{Re}\, q \leqslant -a,\ |\mathrm{Im}\, q| < \theta\},$$
$$R_3 = \{q \mid |\mathrm{Re}\, q| < a,\ |\mathrm{Im}\, q| < \theta\}.$$

定义 2.1 公式(2.2)称作是刚性稳定的,如果它是收敛的,并且对某正数 D, θ, a 有 $R_1 \cup R_2 \subset R$ 和 $R_3 \subset R_s$, 其中 R 和 R_s 分别是公式(2.2)的稳定区域和相对稳定区域.

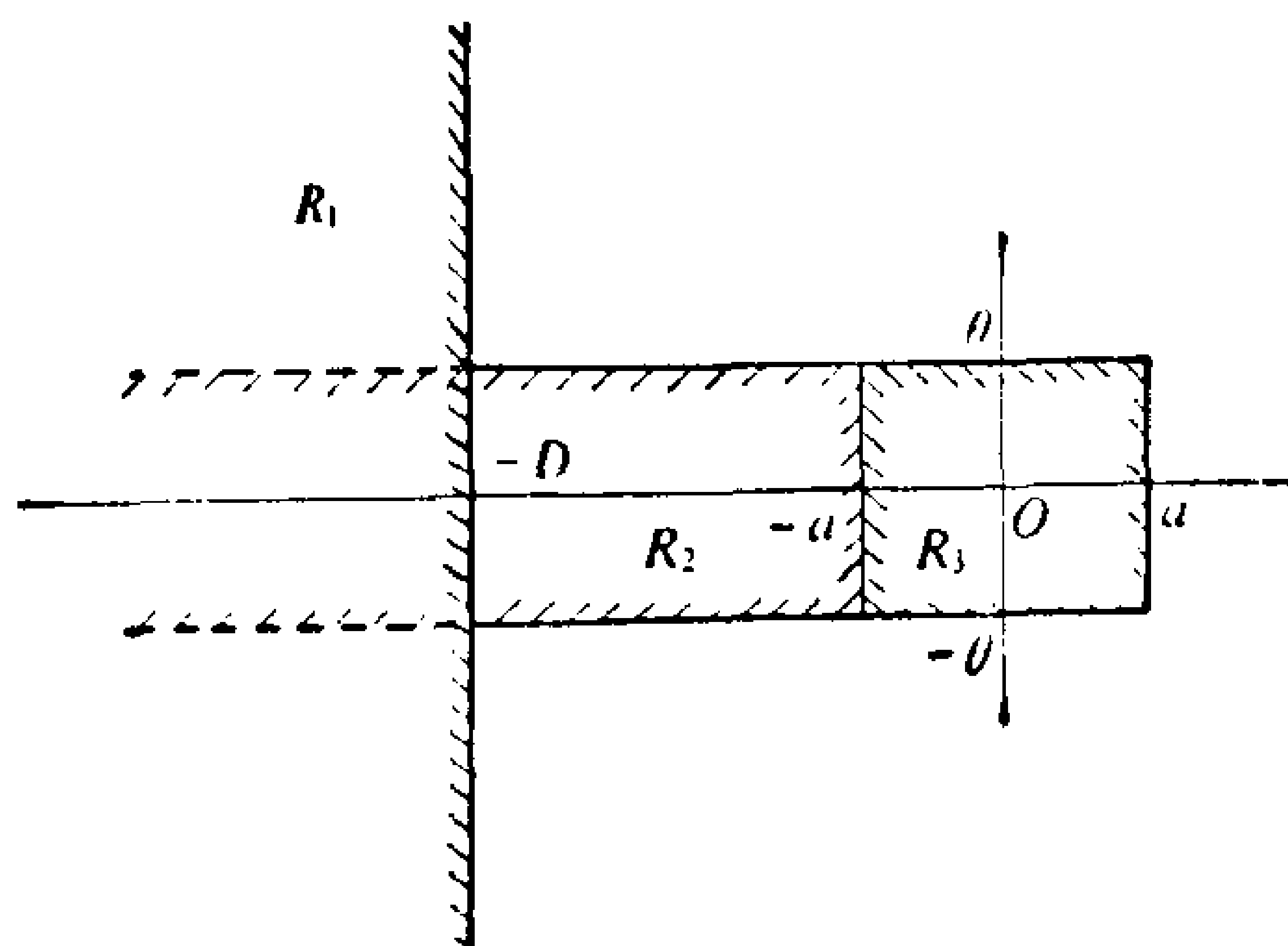


图 2.3 刚性稳定性区域

下面的定理给出方法为刚性稳定的判别准则, 它仅利用多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 的零点条件.

定理 2.15 收敛的公式(2.2)为刚性稳定的充分必要条件是(i)—(iv)成立.

(i) 公式(2.2)是 A_0 稳定的.

(ii) 多项式$\rho(\zeta)/(\zeta-1)$的任何根的模均小于 1.

(iii) 模为 1 的 $\sigma(\zeta)$ 的根是单根.

(iv) 令 ζ 是 $\sigma(\zeta)$ 的模为 1 的根,则 $\rho(\zeta)/(\zeta\sigma'(\zeta))$ 是实数,并且是正的.

证明　必要性　条件 (ii) 的必要性由定义 2.1 立即推得. 可以用证明定理 2.15 的 (ii) 的类似的方法来证明条件 (iii). 为了证明条件 (i), 只要证明由 $[-a,0]$ 上的 $q\in R_s$ 可以推出 $q\in R$. 考虑变换(2.17)及相应的多项式 $r(z)$ 和 $s(z)$. 方程(2.10)变换成

$$r(z)-qs(z)=0,$$

由于公式是收敛的和刚性稳定的, $r(z)$ 和 $s(z)$ 只具有 $\mathrm{Re}\ z\leqslant 0$ 的根, 并且有 $a_{k-1}=2b_k$. 不失一般性, 可以假定有 $a_i\geqslant 0$, $b_i\geqslant 0$. 于是对于任意 $q\in(-a,0)$, $p(z)=r(z)-qs(z)$ 是具有非负系数的多项式. 现在证明它的零点在左开平面内. 首先有 $p(0)=a_0-qb_0\neq 0$, 否则 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 将具有公因子. 另外, $p(z)$ 无正的零点. 现在令 z 是 $p(z)$ 的零点, 且有 $\mathrm{Re}\ z\geqslant 0$, $\mathrm{Im} z>0$. 由于 $p(z)$ 是具实系数的多项式, 因而 $\bar{z}$ 也是其零点. 由变换(2.17), $\zeta=\dfrac{z+1}{z-1}$ 和 $\bar{\zeta}$ 是(2.10)的两个具有同样模的零点, 并且模均大于或等于 1. 这时主根 $\zeta_1(q)$ 所对应的 z 应该是 $p(z)$ 的正根. 这与上面无正根的断言相矛盾. 因此对于 $q\in(-a,0)$, 多项式 $p(z)$ 的根均在左开半平面内. 这样, 对于 $q\in(-a,0)$, 方程 (2.10) 的零点的模均小于 1, 即 $q\in R$. 这表示刚性稳定的方法是 A_0 稳定的. 若 $\bar{\zeta}$ 是多项式 $\sigma(\zeta)$ 的模为 1 的根, 类似于证明定理 2.14 的条件 (iv) 的方法, 我们可得表示式

$$\frac{\rho(\bar{\zeta})}{\bar{\zeta}\sigma'(\bar{\zeta})}=de^{i\varphi},$$

其中 $d>0$, $\varphi\in\left(-\dfrac{\pi}{2},\dfrac{\pi}{2}\right)$. 若 $\varphi\neq 0$, 则我们总可以选取 $\psi\in\left(\dfrac{\pi}{2},\dfrac{3\pi}{2}\right)$, 使得 $\cos(\varphi+\psi)>0$. 当取 $\dfrac{1}{q}=\mu e^{i\psi}$, 和正数 μ 充分小时,

$$\zeta(q)=\bar{\zeta}\left(1+\frac{\rho(\bar{\zeta})}{\bar{\zeta}\sigma'(\bar{\zeta})}\cdot\frac{1}{q}+O\left(\left|\frac{1}{q}\right|^2\right)\right)$$

的模将大于 1. 所以由方法在 R_1 中的性质, 推出 $\varphi=0$. 这就是条件 (iv).

充分性　设收敛的线性多步公式的多项式 $\rho(\zeta)$ 和 $\sigma(\zeta)$ 满足条件 (i)—(iv)．于是由定理 2.14，方法是 $A(0)$ 稳定的．显然，由条件 (ii) 推出，存在一个以 $q=0$ 为心的圆盘属于这方法的相对稳定性区域．因此存在正数 a 和 θ，使得由它们定义的集合 R_2 和 R_3 满足

$$R_2 \subset R,\quad R_3 \subset R_s.$$

为完成定理的充分性证明，仅须证明存在正数 D，使当 $|\zeta|>1$ 时，有 $\mathrm{Re}\,q(\zeta)=\mathrm{Re}(\rho(\zeta)/\sigma(\zeta))\geqslant -D$．令 $\zeta_1, \zeta_2, \cdots, \zeta_m$ 是 $\sigma(\zeta)$ 的模为 1 的根．由条件 (iii) 可以记

$$q(\zeta)=\frac{\rho(\zeta)}{\sigma(\zeta)}=\frac{\rho(\zeta_1)}{\zeta_1\sigma'(\zeta_1)}\frac{\zeta_1}{(\zeta-\zeta_1)}+\cdots$$

$$+\frac{\rho(\zeta_m)}{\zeta_m\sigma'(\zeta_m)}\frac{\zeta_m}{(\zeta-\zeta_m)}+l(\zeta), \tag{2.43}$$

其中 $l(\zeta)$ 是仅在单位圆盘内有极点的有理函数．存在常数 D_l，使在单位圆上有估计 $|l(\zeta)|<D_l$．因此，只要 $|\zeta|\geqslant 1$，一定有 $\mathrm{Re}\,l(\zeta)\geqslant -D_l$．我们看到对于(2.43)中其它的项和所有 $|\zeta|>1$ 有

$$\mathrm{Re}\left[\frac{\rho(\zeta_i)}{\zeta_i\sigma'(\zeta_i)}\frac{\zeta_i}{\zeta-\zeta_i}\right]>-\frac{1}{2}\frac{\rho(\zeta_i)}{\zeta_i\sigma'(\zeta_i)},$$

选取

$$D=\frac{1}{2}\sum_{i=1}^{m}\frac{\rho(\zeta_i)}{\zeta_i\sigma'(\zeta_i)}+D_l$$

将能满足我们的要求．定理证毕

注 2.4　由定理 2.15 的证明可知，刚性稳定的方法一定是 $A(0)$ 稳定的．但反之不然．多项式

$$\rho(\zeta)=\zeta^3-\frac{1}{2}\zeta^2-\frac{1}{2},$$

$$\sigma(\zeta)=\frac{3}{2}\zeta^3-\zeta^2+\frac{3}{2}\zeta$$

所对应的方法是收敛的和 $A(0)$ 稳定的，但它不是刚性稳定的．

例 2.8　当 $d\geqslant 2^{k+1}$ 时，对应于多项式(2.38)的线性的 k 阶 k

步方法是刚性稳定的．因此对于任意阶 p，总存在 p 阶的刚性稳定的方法．

下面来证明对于任意给定的正常数D和任意的正整数 k，总可以找到满足 $d \geqslant 2^{k+1}$ 的数 d，使相应的 Cryer 方法是刚性稳定的 k 步 k 阶方法，并且有 $R_1 = \{q | \operatorname{Re} q < -D\} \subset R$．

设D和k给定，考虑 $d \geqslant 2^{k+1}$ 的 Cryer 方法．为证明上述断言，只须证明总可选取参数 d，使对所有 $\varphi \in [0, 2\pi]$ 有

$$\operatorname{Re} q(e^{i\varphi}) = \operatorname{Re}(\rho(e^{i\varphi})/\sigma(e^{i\varphi})) > -D.$$

为此，用 $r(z)$ 和 $s(z)$ 来表示 q，作一些运算后，得到

$$q(z) = \frac{r(z)}{s(z)} = \frac{r(dw)}{s(dw)}$$

$$= \frac{T_1(w) + 2[(1+w)^k - 1]/w}{d(1+w)^k},$$

其中 $z = dw$，$T_1(w)$ 由例 2.5 中给出．由于变换

$$\zeta \to z = (\zeta + 1)/(\zeta - 1)$$

将单位圆映射到虚轴上，我们必须考虑 $q(iy)$，y 为实数．在例 2.5 中已指出对所有的实数 y 和所有的 $d \geqslant 2^{k+1}$ 有估计

$$|T_1(iy)| \leqslant \frac{1}{12} \max(1, y^{k-1}).$$

因此有

$$\frac{|T_1(iy)|}{|1 + iy|^k} \leqslant \frac{1}{12},$$

有理函数 $[(1+w)^k - 1]/[w(1+w)^k]$ 在无穷远处是有界的，并且在虚轴 $\operatorname{Re} w = 0$ 上无极点．因此存在常数 C，使对所有实数 y 有估计

$$\left| 2\frac{(1+iy)^k - 1}{(1+iy)^k iy} \right| < C.$$

选取

$$d = \max\left\{2^{k+1}, \left(C + \frac{1}{12}\right)\Big/D\right\},$$

则对所有实数 y 有

$$|q(iy)| < \frac{1}{d}\left(\frac{1}{12} + C\right) \leqslant D,$$

因此有 $\operatorname{Re} q(iy) > -D$. 这就是要证明的.

在证明中，我们得到 Cryer 方法的稳定区域含有圆盘

$$\left\{q \,\middle|\, |q| < \frac{1}{d}\left(\frac{1}{12} + C\right)\right\}$$

的整个外部. 在图 2.4 中画出了 $k = 1, 2, \cdots, 7, d = 2^{k+1}$ 的 Cryer 方法的稳定区域. 这些区域是在复平面上通过描出曲线 $q = \rho(e^{i\varphi})/\sigma(e^{i\varphi}), \varphi \in [0, 2\pi]$ 得到的. 由于区域对实轴是对称的. 我们省掉下半部分.

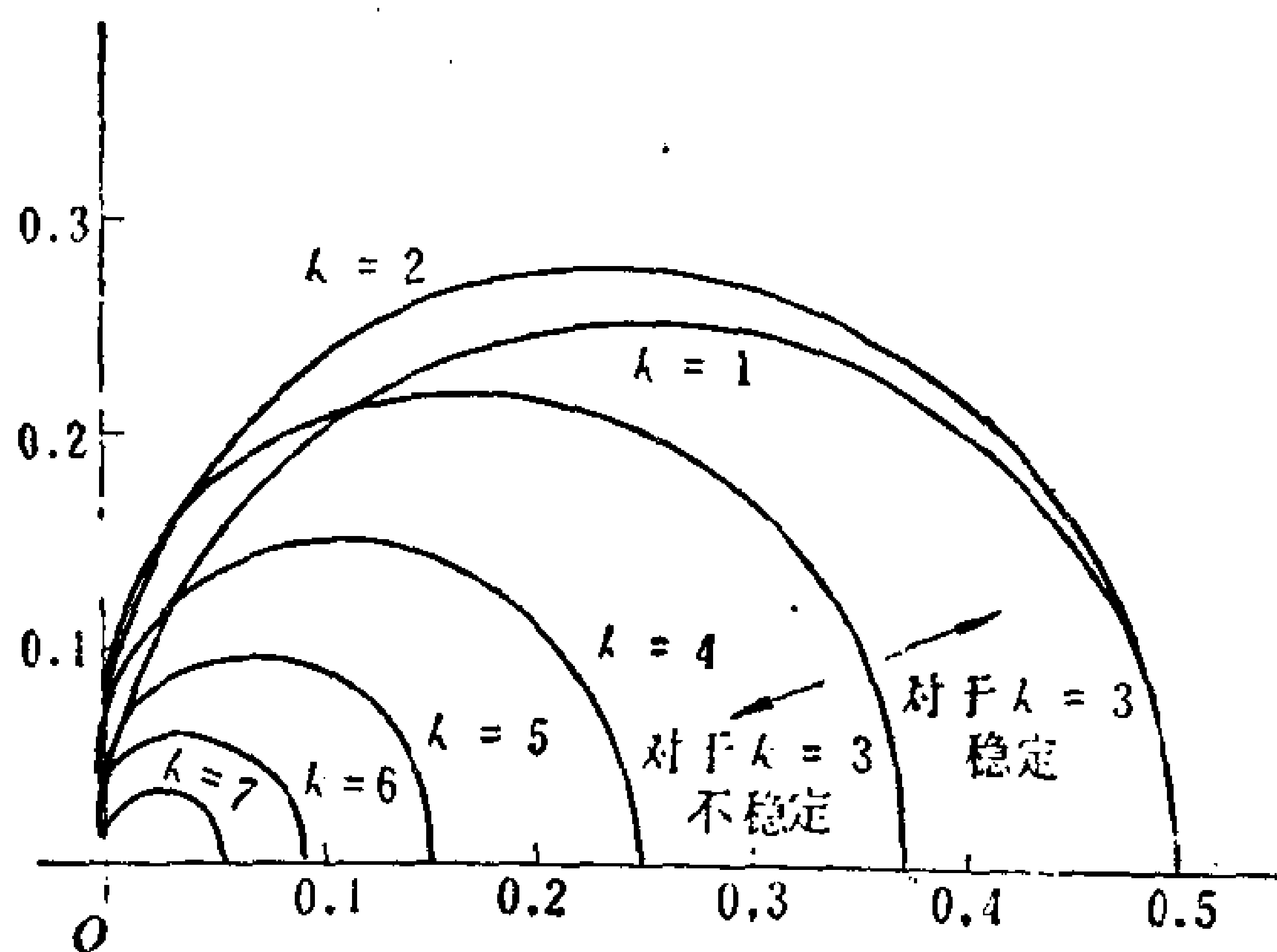


图 2.4 Cryer k 步方法的稳定区域

$k = 1, 2, \cdots, 7, d = 2^{k+1}$, 在圆的外部是稳定的

在表 2.2(a)中列出 Cryer 方法的 $A(\alpha)$稳定性和刚性稳定性的参数 D 和 α, 其中参数 D 的括号中的数表示括号前的数应乘上 10 的幂次. 表 2.2(b)说明当 i 增加时, α 收敛到 83.59561833°, 而 D 收敛到零. 这与上面的证明结果是一致的. 应该指出, 当 k 或 i 增大时, $\rho(\zeta)$ 的系数的绝对值变得非常小, 而 $\sigma(\zeta)$ 的系数变得非常大, 它们之间的相对量级差得很多.

表 2.2 Cryer 方法的 $A(\alpha)$ 和刚性稳定性的参数 α 和 D

(a) $d = 2^{k+1}, k = 1, 2, \cdots, 16$

k	D	α
1	A稳定	A稳定
2	A稳定	A稳定
3	1.17(−3)	88.8°
4	3.31(−3)	86.3°
5	4.03(−3)	83.6°
6	3.60(−3)	81.0°
7	2.75(−3)	78.5°
8	1.90(−3)	76.3°
9	1.24(−3)	74.2°
10	7.74(−4)	72.3°
11	4.68(−4)	70.5°
12	2.76(−4)	68.9°
13	1.59(−4)	67.3°
14	9.07(−5)	65.9°
15	5.10(−5)	64.6°
16	2.84(−5)	63.3°

(b) $k = 5, d = 2^{k+i}, i = 1, 2, \cdots, 10$

i	D	α
1	4.03(−3)	83.58858608°
2	2.01(−3)	83.59386044°
3	1.01(−3)	83.59517887°
4	5.03(−4)	83.59550847°
5	2.52(−4)	83.59559087°
6	1.26(−4)	83.59561147°
7	6.29(−5)	83.59561662°
8	3.15(−5)	83.59561791°
9	1.57(−5)	83.59561829°
10	7.87(−6)	83.59561831°

本 章 附 注

§2 的材料主要取自 Dahlquist [48], Liniger[77] 和 Genin[59].

§3 的材料取自 Widlund[113] 和 Nørsett [89].

§4 的材料取自 Cryer[46] 和 Jeltsch[64].

§5 的材料取自 Jeltsch[64].

第三章　向后差分方法

向后差分的数值积分公式

$$\sum_{i=1}^{k} \frac{1}{i} \nabla^i y_{n+k} = hf_{n+k} \tag{3.1}$$

可以用来计算微分方程初值问题的数值解是早就知道的，但是由于它仅当$0 < k \leqslant 6$时才满足零稳定性条件，并且精度阶比通常的 Adams-Moulton 方法要差一点(公式(3.1)的精度阶是 k，而k步的 Adams-Moulton 方法的精度阶是$k+1$)，所以在实际中几乎没有使用公式(3.1)．直到1968年，Gear 指出它在无穷远处具有好的稳定性质，即具有他所提出的刚性稳定性，并且以公式(3.1)为基础编制了一个求解刚性微分方程组的通用程序包．从这以后，公式(3.1)才得到广泛的应用和详尽的研究．据文献[55]中的对求解刚性方程的许多方法的比较，向后差分公式(3.1)是效果最好的几个方法之一．十多年来，虽然已提出过许多有效的算法，但是它仍然是最有效的通用算法．究其原因，主要是它具有下面的三个优点：1)容易改变阶和步长，2)能够应用高阶的和高稳定的格式，3)每前进一个步长解隐式方程组所需要的工作量比较小．显然一个算法若要具有与 Gear 算法可比较的效果，至少必须具有上述的三点．但是从目前的方法来看，同时满足上面的三点的方法是较难找到的．

§1　向后差分公式

对于数值积分初值问题

$$\frac{dy}{dt} = f(t, y), \quad y(0) = y_0 \tag{3.2}$$

的差分方法通常是先将问题(3.2)的微分方程化成

$$y(t_{n+k}) - y(t_n) = \int_{t_n}^{t_{n+k}} f(t, y(t))\, dt .$$

然后用插值多项式代替被积函数，再通过积分插值公式的基函数得到差分公式. 另外，也可以直接用插值多项式来近似 $y(t)$，然后再通过微分得到 y'，代替 (3.2) 中微分方程左端以导出差分公式. 这一节讨论用后一种方法.

以 $t_n, t_{n+1}, \cdots, t_{n+k}$ 作为插值节点的函数 $y(t)$ 的插值多项式 $p(t)$ 为

$$P(t) = \sum_{m=0}^{k} (-1)^m \binom{-s}{m} \nabla^m y_{n+k}, \tag{3.3}$$

其中

$$\binom{q}{0} = 1,$$

$$\binom{q}{m} = \frac{q(q-1)\cdots(q-m+1)}{m!}, \quad m = 1, 2, \cdots,$$

$$s = \frac{t - t_{n+k}}{h}.$$

在向后差分 $\nabla^m y_{n+k}$ 中 y_α 是 $y(t)$在节点 t_α 处的近似值. 在点 $t = t_{n+i}$ 处对函数 $p(t)$ 求导得

$$P'(t_{n+i}) = \frac{1}{h} \sum_{m=0}^{k} \delta_{i,m} \nabla^m y_{n+k}, \tag{3.4}$$

其中

$$\begin{aligned} \delta_{i,m} &= (-1)^m h \frac{d}{dt} \binom{-s}{m}_{t=t_{n+i}} \\ &= (-1)^m \frac{d}{ds} \binom{-s}{m}_{s=-(k-i)}. \end{aligned} \tag{3.5}$$

令 $P'(t_{n+i}) = f(t_{n+i}, y_{n+i})$，我们得到方程(3.2)的差分近似

$$\sum_{m=0}^{k} \delta_{j,m} \nabla^m y_{n+k} = h f_{n+j}. \tag{3.6}$$

现在来计算 $\delta_{j,m}$. 如果 $m > k - j$, 应用二项式系数的定义

$$\delta_{j,m} = \frac{d}{ds}\left\{\frac{s(s+1)\cdots(s+m-1)}{m!}\right\}_{s=-(k-j)}$$

$$= \frac{1}{m!}\lim_{s\to-(k-j)} s(s+1)\cdots(s+k-j-1)$$

$$\times (s+k-j+1)\cdots(s+m-1)$$

$$= (-1)^{k-j}\frac{(k-j)!(m-k+j-1)!}{m!},$$

对于 $m \leqslant k - j$, 直接微分不太方便,可以应用生成函数来建立。设

$$D_j(t) = \sum_{m=0}^{\infty} \delta_{j,m} t^m,$$

于是我们有

$$D_j(t) = \sum_{m=0}^{\infty} (-t)^m \frac{d}{ds}\binom{-s}{m}_{s=-(k-j)}$$

$$= \frac{d}{ds}\left(\sum_{m=0}^{\infty}\binom{-s}{m}(-t)^m\right)_{s=-(k-j)}$$

$$= \frac{d}{ds}(1-t)^{-s}\Big|_{s=-(k-j)}$$

$$= \frac{d}{ds} e^{-s\log(1-t)}\Big|_{s=-(k-j)}.$$

因此有

$$D_j(t) = -\log(1-t)\cdot(1-t)^{(k-j)}.$$

这就推出

$$\delta_{j,0} + \delta_{j,1}t + \delta_{j,2}t^2 + \cdots = \left(t + \frac{1}{2}t^2 + \frac{1}{3}t^3 + \cdots\right)$$

$$\times\left[1-\binom{k-j}{1}t+\binom{k-j}{2}t^2-\cdots+(-1)^{k-j}t^{k-j}\right].$$

比较同次幂的系数，有 $\delta_{j,0}=0(j=0,1,2,\cdots)$ 和

$$\delta_{j,m}=\frac{1}{m}-\frac{1}{m-1}\binom{k-j}{1}+\frac{1}{m-2}\binom{k-j}{2}-\cdots$$

$$+(-1)^{m-1}\binom{k-j}{m-1},\quad m\leqslant k-j,$$

$$\delta_{j,m}=\frac{1}{m}-\frac{1}{m-1}\binom{k-j}{1}+\frac{1}{m-2}\binom{k-j}{2}-\cdots$$

$$+(-1)^{k-j}\frac{1}{m-k+j},\quad m>k-j.$$

由这公式，我们特别可以得到

$$\delta_{k,m}=\frac{1}{m},\ m\geqslant 1,$$

$$\delta_{k,0}=0,$$

这时公式(3.6)取形式

$$\sum_{m=1}^{k}\frac{1}{m}\nabla^m y_{n+k}=hf_{n+k}. \tag{3.7}$$

为了得到公式(3.6)的局部截断误差，我们引用如下的引理.

引理 3.1 令 J 是含 $k+1$ 个不同的点 $t_\nu(\nu=0,1,\cdots,k)$的最小区间，函数 $z(t)$具有 $k+1$ 阶的连续导数，$P(t)$是以 t_ν 为节点的 $z(t)$的插值多项式. 于是对于每一个 $t_\nu(\nu=0,1,\cdots,k)$，存在数$\xi\in J$ 使得有

$$z'(t_\nu)-P'(t_\nu)=\frac{1}{(k+1)!}z^{(k+1)}(\xi)L'(t_\nu),$$

其中 $L(t)=(t-t_0)(t-t_1)\cdots(t-t_k)$，$\xi\in J$.

由引理 1 可得公式(3.6)的局部截断误差为

$$\frac{h}{(k+1)!}y^{(k+1)}(\xi)L'(t_{n+j}),\quad t_n<\xi<t_{n+k}.$$

它还可以表成

$$h^{k+1}g_{i,k+1}y^{(k+1)}(\xi), \quad t_n < \xi < t_{n+k}.$$

对于公式(3.7)的局部截断误差为

$$\frac{1}{k+1}h^{(k+1)}y^{(k+1)}(\xi), \quad t_n < \xi < t_{n+k}. \tag{3.8}$$

将公式(3.7)写成一般的线性多步公式的形式，有

$$y_{n+k} = \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h\beta_k f_{n+k}. \tag{3.9}$$

对于 $k = 1, \cdots, 6$，公式(3.9)中的系数 α_i，β_k 由表 3.1 给出。

表 3.1　向后差分方法的系数

k	1	2	3	4	5	6
β_k	1	$\frac{2}{3}$	$\frac{6}{11}$	$\frac{12}{25}$	$\frac{60}{137}$	$\frac{60}{147}$
α_0	1	$-\frac{1}{3}$	$\frac{2}{11}$	$-\frac{3}{25}$	$\frac{12}{137}$	$-\frac{10}{147}$
α_1		$\frac{4}{3}$	$-\frac{9}{11}$	$\frac{16}{25}$	$-\frac{75}{137}$	$\frac{72}{147}$
α_2			$\frac{18}{11}$	$-\frac{36}{25}$	$\frac{200}{137}$	$-\frac{225}{147}$
α_3				$\frac{48}{25}$	$-\frac{300}{137}$	$\frac{400}{147}$
α_4					$\frac{300}{137}$	$-\frac{450}{147}$
α_5						$\frac{360}{147}$

公式(3.9)是一个隐式方程，需要用一个显式方程来预估 y_{n+k}。可以将这种预估方程取为

$$y_{n+k} = \sum_{i=0}^{k-1} \alpha_i^* y_{n+i} + h\beta_{k-1}^* f_{n+k-1}, \tag{3.10}$$

选取其中的系数 α_i^*，$i = 0, 1, \cdots, k-1$ 和 β_{k-1}^* 使公式(3.10)的精度阶也是 k。于是可以用公式(3.10)作为预估求出 y_{n+k} 的初值

$y_{n+k}^{(0)}$，再用公式(3.9)来进行校正，求出满足方程

$$y_{n+k} = \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h\beta_k f(t_{n+k}, y_{n+k}) \tag{3.11}$$

的解 y_{n+k}. (3.11)是 y_{n+k} 的非线性方程组. 对于非刚性方程，若用简单迭代法求解(3.11)，即

$$y_{n+k}^{(m+1)} = \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h\beta_k f(t_{n+k}, y_{n+k}^{(m)}), \tag{3.12}$$

其中 $y_{n+k}^{(0)}$ 由方程(3.10)给出. 为了保证迭代(3.12)的收敛性，要求有

$$\left| h\beta_k \frac{\partial f}{\partial y} \right| < 1. \tag{3.13}$$

对于刚性方程，Jacobi 矩阵的模 $\left| \frac{\partial f}{\partial y} \right|$ 很大，(3.13)是对 h 的另一种限制，它类似于第一章 § 1.1 中讨论的计算稳定性对步长 h 的约束. 因此满足(3.13)的 h 必须取得很小. 为了克服这种困难，通常采用 Newton-Raphson 方法或者修改的 Newton 方法. 若采用 Newton-Raphson 方法，迭代格式(3.12)换成

$$y_{n+k}^{(m+1)} = y_{n+k}^{(m)} - w_{n+k}^{(m)} \left[y_{n+k}^{(m)} - \sum_{i=0}^{k-1} \alpha_i y_{n+i} - h\beta_k f(t_{n+k}, y_{n+k}^{(m)}) \right], \tag{3.14}$$

其中

$$w_{n+k}^{(m)} = \left[I - h\beta_k \frac{\partial f}{\partial y}(t_{n+k}, y_{n+k}^{(m)}) \right]^{-1}. \tag{3.15}$$

将 $w_{n+k}^{(m)}$ 用某一个固定的量来代替，就得到修改的 Newton 方法.

Gear 以公式(3.10)和(3.11)为基础编制了一个求解刚性方程组的程序包[20]. 这个程序包具有自动变阶，变步长和选取计算起始值的特点. 这就是所谓的自动积分，即只要给出微分方程组的初始值，程序就能自动地计算出使用公式(3.10)和(3.11)所需要的起始值，并且在数值积分的过程中，根据给定的精度要求改变积分

步长和积分公式的阶，使计算工作量尽可能小。下面我们对程序包中所用的一些处理作一简单的介绍。

首先 Gear 为了变步长和变阶的方便，将方法表示成 Nordsieck 形式。Nordsieck 提出贮存 y_n，hy'_n，$\frac{h^2}{2!}y''_n\cdots$ 的近似值来代替贮存 $y(t)$及其导数在各个节点上的近似值。他的出发点是使改变步长所需要的附带计算简单一点。将方程（3.2）的解 $y(t)$及其各阶导数展成 Taylor 级数，并乘以因子 $\frac{h^i}{i!}$，我们得到

$$\begin{aligned}
y(t+h) &= y(t) + \frac{h}{1!}y'(t) + \frac{h^2}{2!}y''(t) \\
&\qquad + \frac{h^3}{3!}y^{(III)}(t) + \frac{h^4}{4!}y^{(IV)}(t) + \cdots, \\
\frac{h}{1!}y'(t+h) &= \frac{h}{1!}y'(t) + 2\cdot\frac{h^2}{2!}y''(t) \\
&\qquad + 3\cdot\frac{h^3}{3!}y^{(III)}(t) + 4\cdot\frac{h^4}{4!}y^{(IV)}(t) + \cdots, \\
\frac{h^2}{2!}y''(t+h) &= \frac{h^2}{2!}y''(t) + 3\cdot\frac{h^3}{3!}y^{(III)}(t) \\
&\qquad + 6\cdot\frac{h^4}{4!}y^{(IV)}(t) + \cdots, \qquad (3.16) \\
&\cdots\cdots
\end{aligned}$$

记

$$z_n = \left[y_n,\ \frac{h}{1!}y'_n,\ \frac{h^2}{2!}y''_n,\cdots,\frac{h^k}{k!}y_n^{(k)}\right], \qquad (3.17)$$

其中 y_n，y'_n，$\cdots$，是 $y(t_n)$，$y'(t_n)$，$\cdots$ 的近似值。如果公式中所用的导数多，表示式(3.17)中可以含有更高阶的导数的信息。由(3.16)，z_{n+1} 可由 z_n 来预估

$$z_{n+1} = Az_n, \qquad (3.18)$$

其中矩阵A是 Pascal 上三角阵. 它的第 (i, j) 元素是

$$\binom{j-1}{i-1} \ (j \geqslant i),$$

或者为零$(j < i)$. 例如对于 6×6 Pascal 矩阵,有

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ & 1 & 2 & 3 & 4 & 5 & 6 \\ & & 1 & 3 & 6 & 10 & 15 \\ & & & 1 & 4 & 10 & 20 \\ & & & & 1 & 5 & 15 \\ & & & & & 1 & 6 \\ & & & & & & 1 \end{bmatrix}.$$

容易看到,公式(3.18)未用到微分方程, 从数值计算来看它是不稳定的. 为了克服这种不稳定性,再加上一个校正迭代,而把公式(3.18)仅作为零次近似. 完整的公式为

$$\begin{aligned} z_{n+1}^{(0)} &= A z_n, \\ z_{n+1}^{(m+1)} &= z_{n+1}^{(m)} + l w_m F(z_{n+1}^{(m)}). \end{aligned} \tag{3.19}$$

这就是公式(3.10)、(3.14)的 Nordsieck 的矩阵表示形式, 其中量 l, w, F 的具体形式在下面确定.

为了由公式(3.10)、(3.11)、(3.14)来确定(3.19)中的未确定的量,我们将公式(3.10)、(3.11)、(3.14)表示成矩阵形式.

公式(3.10)、(3.11)中所需要贮存的信息是 y_{n+k}, y_{n+k-1}, $\cdots$, y_n, hy'_{n+k}, 令

$$Y_{n+k} = [y_{n+k}, hy'_{n+k}, y_{n+k-1}, \cdots, y_n]^T, \tag{3.20}$$

由(3.10)式,并将(3.11)式左端的 y_{n+k} 代之以(3.10)式右端,解出 $hy'_{n+k} = hf(t_{n+k}, y_{n+k})$, 推得预估式

$$Y_{n+k}^{(0)} = BY_{n+k-1}, \tag{3.21}$$

其中

$$
B=\begin{bmatrix}
\alpha_{k-1}^{*} & \beta_{k-1}^{*} & \alpha_{k-2}^{*} & \cdots & \alpha_{0}^{*} \\
\dfrac{\alpha_{k-1}^{*}-\alpha_{k-1}}{\beta_k} & \dfrac{\beta_{k-1}^{*}}{\beta_k} & \dfrac{\alpha_{k-2}^{*}-\alpha_{k-2}}{\beta_k} & \cdots & \dfrac{\alpha_{0}^{*}-\alpha_0}{\beta_k} \\
1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
 & & \cdots\cdots & & \\
0 & 0 & 0 & \cdots & 1\ 0
\end{bmatrix}. \tag{3.22}
$$

若用简单迭代法，可推得下面的迭代格式

$$
\begin{cases}
y_{n+k}^{(1)}=y_{n+k}^{(0)}-\beta_k[hy_{n+k}^{(0)\prime}-hf(t_{n+k},y_{n+k}^{(0)})],\\
hy_{n+k}^{(1)\prime}=hy_{n+k}^{(0)\prime}+[-hy_{n+k}^{(0)\prime}+hf(t_{n+k},y_{n+k}^{(0)})]\\
\qquad =y_{n+k}^{(1)}-\sum\limits_{i=0}^{k-1}\alpha_i y_{n+i},\\
y_{n+k-i}^{(1)}=y_{n+k-i},\quad i=1,2,\cdots,k.
\end{cases} \tag{3.23}
$$

一般可得

$$
Y_{n+k}^{(m+1)}=Y_{n+k}^{(m)}+\tilde{l}F(Y_{n+k}^{(m)}), \tag{3.24}
$$

其中

$$
\tilde{l}=(+\beta_k,+1,0,\cdots,0), \tag{3.25}
$$

$$
F(Y_{n+k}^{(m)})=-hy_{n+k}^{(m)\prime}+hf(t_{n+k},y_{n+k}^{(m)}). \tag{3.26}
$$

记号 $\tilde{l}F$ 表示一个向量，其第 i 个分量为 $\tilde{l}$ 的第 i 个分量乘上 F. 为了将公式(3.21)、(3.24)转换成用基向量 z_n 来表示，只要求出由 Y_n 到 z_n 的转换矩阵 Q，使

$$
z_n=QY_n.
$$

把(3.21)和(3.24)的两端乘上 Q，得到

$$
z_{n+k}^{(0)}=QY_{n+k}^{(0)}=QBY_{n+k-1}=QBQ^{-1}z_{n+k-1}, \tag{3.27}
$$

$$
\begin{aligned}
z_{n+k}^{(m+1)}&=QY_{n+k}^{(m+1)}=QY_{n+k}^{(m)}+Q\tilde{l}F(Q^{-1}z_{n+k}^{(m)})\\
&=z_{n+k}^{(m)}+Q\tilde{l}F(Q^{-1}z_{n+k}^{(m)}),
\end{aligned} \tag{3.28}
$$

转换矩阵 Q 可以由基向量 Y_n 和 z_n 对所有次数不超过 k 次的多项式的等价性来构成. 由于 $F(Y_{n+k}^{(m)})$ 仅依赖于 $hy_{n+k}^{\prime(m)}$ 和 $y_{n+k}^{(m)}$，所以推得 $F(Y_{n+k}^{(m)})=F(Q^{-1}z_{n+k}^{(m)})=F(z_{n+k}^{(m)})$. 由于(3.27)和(3.18)对所有次数不超过 k 次的多项式精确成立，可以推得

$$QBQ^{-1} = A,$$

令 $l = Q\bar{l} = [l_0, l_1, \cdots, l_k]^T$，则迭代(3.27)、(3.28)可以记成

$$\begin{aligned} z_{n+k}^{(0)} &= A z_{n+k-1}, \\ z_{n+k}^{(m+1)} &= z_{n+k}^{(m)} + lF(z_{n+k}^{(m)}). \end{aligned} \tag{3.29}$$

象迭代格式(3.12)一样，对于刚性方程，格式(3.29)的收敛性要求步长 h 非常小，需要采用 Newton 方法来求解。如果(3.29)收敛，$z_{n+k}^{(m)}$ 将收敛到

$$z_{n+k} = z_{n+k}^{(0)} + lw, \tag{3.30}$$

其中 w 满足

$$F(z_{n+k}) = F(z_{n+k}^{(0)} + lw) = 0. \tag{3.31}$$

应用 Newton 方法求方程(3.31)的解 w，我们将得到迭代

$$w_{(m+1)} = w_{(m)} - \left[\frac{\partial F}{\partial z} \cdot l\right]^{-1} F(z_{n+k}^{(0)} + lw_{(m)}). \tag{3.32}$$

如果记 $z_{n+k}^{(m)} = z_{n+k}^{(0)} + lw_{(m)}$，(3.32) 式变成

$$z_{n+k}^{(m+1)} = z_{n+k}^{(m)} - l\left[\frac{\partial F}{\partial z} \cdot l\right]^{-1} F(z_{n+k}^{(m)}). \tag{3.33}$$

由于 $F(z_{n+k}) = -hy'_{n+k} + hf(t_{n+k}, y_{n+k})$，有

$$w = \left[\frac{\partial F}{\partial z} \cdot l\right]^{-1} = \left[-l_1 I + hl_0 \frac{\partial f}{\partial y}\right]^{-1}, \tag{3.34}$$

其中 $l_0 = +\beta_k, l_1 = +1$。我们看到 w 依赖于方法的阶(通过 β_k)，h 和 $\dfrac{\partial f}{\partial y}$。如果 $\partial f/\partial y$ 是慢变的(在实际中经常发生)，则对一步或其中步长和阶不变的若干步，在迭代(3.33)过程中 w 将变化不大。在 Gear 的程序中利用了这个事实。只有在第三次迭代误差仍不小的意义下，当改变阶或校正量 $WF(z_{n+k}^{(m)})$，校正过程不收敛时，矩阵 w 才重新计算。

在表 3.2 中列出 l 的各个分量，它相应于表 3.1 中列出的系数的方法。

利用基向量 z_n 实现程序的变步长和变阶比较容易。原因是利用基向量 z_n 容易进行误差估计和公式的起步。下面我们叙述

表 3.2 方法(3.33)的向量 l 的分量

l \ k	1	2	3	4	5	6
l_0	1	$\frac{2}{3}$	$\frac{6}{11}$	$\frac{24}{50}$	$\frac{120}{274}$	$\frac{720}{1764}$
l_1	1	$\frac{3}{3}$	$\frac{11}{11}$	$\frac{50}{50}$	$\frac{274}{274}$	$\frac{1764}{1764}$
l_2		$\frac{1}{3}$	$\frac{6}{11}$	$\frac{35}{50}$	$\frac{225}{274}$	$\frac{1624}{1764}$
l_3			$\frac{1}{11}$	$\frac{10}{50}$	$\frac{85}{274}$	$\frac{735}{1764}$
l_4				$\frac{1}{50}$	$\frac{15}{274}$	$\frac{175}{1764}$
l_5					$\frac{1}{274}$	$\frac{21}{1764}$
l_6						$\frac{1}{1764}$

Gear 的程序中的一些处理.

对于 k 阶方法，在计算过程中，解向量 z_n 的最后一个分量 $z_n^{(k)}$ 是 $\frac{h^k}{k!}y_n^{(k)}$. 作向后差分

$$\nabla z_n^{(k)} = \frac{h^k}{k!}(y_n^{(k)} - y_{n-1}^{(k)}), \tag{3.35}$$

则 $\nabla z_n^{(k)}$ 给出量 $h^{k+1}y_n^{(k+1)}/k!$ 的一种估计. 可以从两个方面来应用量 $\nabla z_n^{(k)}$. 一方面当积分方法改成 $k+1$ 阶方法时，可以将 $\frac{1}{k+1}\nabla z_n^{(k)}$ 作为是 $\frac{h^{k+1}}{(k+1)!}y_n^{(k+1)}$ 加到 z_n 中而成为它的第 $k+2$ 个分量. 另一方面可以利用它来进行误差估计.

我们知道,对于 k 阶方法,每步的局部截断误差是

$$c_{k+1}h^{k+1}y_n^{(k+1)} + O(h^{k+2}), \tag{3.36}$$

其中 c_{k+1} 依赖于所使用的方法. 对于公式(3.7),由(3.8),得

$$c_{k+1} = \frac{1}{k+1}. \tag{3.37}$$

我们略去(3.36)中的第二项，只估计第一项。利用量 $\nabla z_n^{(k)}$，这个误差可近似地表示成

$$c_{k+1}h^{k+1}y_n^{(k+1)} \approx c_{k+1}k!\nabla z_n^{(k)}.$$

因此，若计算过程中要求每步的局部截断误差小于预先指定的误差量 ε，必须选取步长 h，使得有

$$c_{k+1}k!\,|\nabla z_n^{(k)}| \leqslant \varepsilon. \tag{3.38}$$

当积分一个方程组时，希望控制每个分量的误差，对于选定的权向量 w，要求选取步长 h 使得不等式

$$c_{k+1}k!\left\|\frac{\nabla z_n^{(k)}}{w}\right\|_2 \leqslant \varepsilon \tag{3.39}$$

成立，其中方程组的每个元有一个 $\nabla z_n^{(k)}$ 分量和权分量. $\|\cdot\|_2$ 是 L_2 范数. 使用 L_2 范数是因为在计算机上计算可快一些. 也可以使用最大模范数来计算.

在 Gear 程序中基本步长的控制程序是积分一步并且检验(3.39)是否成立. 如果(3.39)成立，则接受这一步，否则就抛弃这一步. 对于下一步或重新计算抛弃的步，所用的步长估计为 αh，其中 α 由等式

$$c_{k+1}k!\alpha^{k+1}\left\|\frac{\nabla z_n^{(k)}}{w}\right\|_2 = \varepsilon$$

来确定. 如果采用这个步长，并且误差正好又与 h^{k+1} 成比例(即 $\nabla z_n^{(k)}$ 不变)，则下次检验(3.39)时，(3.39)正好满足. 但是 $\nabla z_n^{(k)}$ 不总是不变的. 所以为了保证不等式(3.39)能够满足，常采用稍小一点的步长. 在程序中 α 由

$$\alpha_k = \frac{1}{1\cdot 2}\left[\frac{\varepsilon}{c_{k+1}k!}\ \frac{1}{\left\|\frac{\nabla z_n^{(k)}}{w}\right\|_2}\right]^{1/k+1}$$

来确定. 为了变阶，还必须检查在其它阶的公式中所用的步长. 由于

$$\nabla^2 z_n^{(k)} \approx \frac{h^{k+2}}{k!}\,y_n^{(k+2)},$$

$$z_n^{(k)} \approx \frac{h^k}{k!} y_n^{(k)},$$

则若在 $k+1$ 阶和 $k-1$ 阶公式中可用的步长记成 αh、α 可估计为

$$\alpha_{k+1} = \frac{1}{1 \cdot 4}\left[\frac{\varepsilon}{c_{k+2}k!} \frac{1}{\left\|\frac{\nabla^2 z_n^{(k)}}{w}\right\|_2}\right]^{1/k+2}, \text{对于 } k+1 \text{ 阶公式,}$$

$$\alpha_{k-1} = \frac{1}{1 \cdot 3}\left[\frac{\varepsilon}{c_k k!} \frac{1}{\left\|\frac{z_n^{(k)}}{w}\right\|_2}\right]^{1/k}, \text{ 对于 } k-1 \text{ 阶公式,}$$

其中因子 $1 \cdot 4$ 和 $1 \cdot 3$ 的选取是这样考虑的：一方面要尽量保持原来的公式使不等式(3.39)成立。另一方面尽量不改变阶，因为变阶时需要额外的计算时间，在需要变阶时尽量利用降阶，因为它每步所用的工作量稍少一点.

在积分过程中，每一步都计算 α_k, α_{k-1} 和 α_{k+1}, 选取其中的最大的 α, 以确定下一步采用的合理的步长和阶. 但是在具体程序中，每步改变步长和阶不一定好，Gear 作了下面的考虑：

1. 如果这一步失败，则估计 α.

2. 在上一次改变阶或步长以后的 $k+1$ 步，估计 α(使阶或步长的改变不频繁).

3. 如果在上次估计 α 时步长不放大，则在其 10 步后估计 α(这可以减少过于频繁的附加计算).

积分时，从一阶公式开始，取 $z_0 = (y_0, hy_0')^T$. 其中 y_0 是已知的，而 $hy_0' = hf(0, y_0)$. 因此程序的自开始是不成问题的. 步长和阶的控制程序能把步长和阶增大到符合要求的程度. 在降阶时，取 α_{k-1}, 去掉基向量中最后一个分量. 当提高一阶时，取 α_{k+1}, 将基向量的最后一个分量的向后差分除以 $k+1$, 并且把它加到基向量中去作为最后一个分量，使方法变成 $k+1$ 阶的.

在[55]中，对求解刚性方程的一些数值方法进行了严格的数值比较，证明 Gear 的程序一般是很有效的，只当方程(3.2)的右函

数的 Jacobi 矩阵的某些特征值接近于虚轴时，Gear 的程序才出现困难，产生不稳定现象．出现这种现象的原因是由于(3.7)的高阶公式的稳定性区域包含虚轴的邻近部分比较小引起的(见 §2)．为了克服这个缺点，许多作者研究了探测并克服不稳定性的技巧，来改进 Gear 的程序．

§2 向后差分公式的稳定性

这一节讨论向后差分公式的稳定性．k 步向后差分公式(3.7)可以由多项式

$$\rho(\xi) = \xi^k \sum_{j=1}^{k} \frac{1}{j}(1-\xi^{-1})^j, \tag{3.40}$$

$$\sigma(\xi) = \xi^k \tag{3.41}$$

来确定．应用于试验方程

$$y' = \lambda y,\ y(0) = 1, \tag{3.42}$$

得到特征方程

$$\rho(\xi) - q\sigma(\xi) = 0,\ q = \lambda h. \tag{3.43}$$

于是通过

$$q = \frac{\rho(\xi)}{\sigma(\xi)} = \sum_{j=1}^{k} \frac{1}{j}(1-\xi^{-1})^j, \tag{3.44}$$

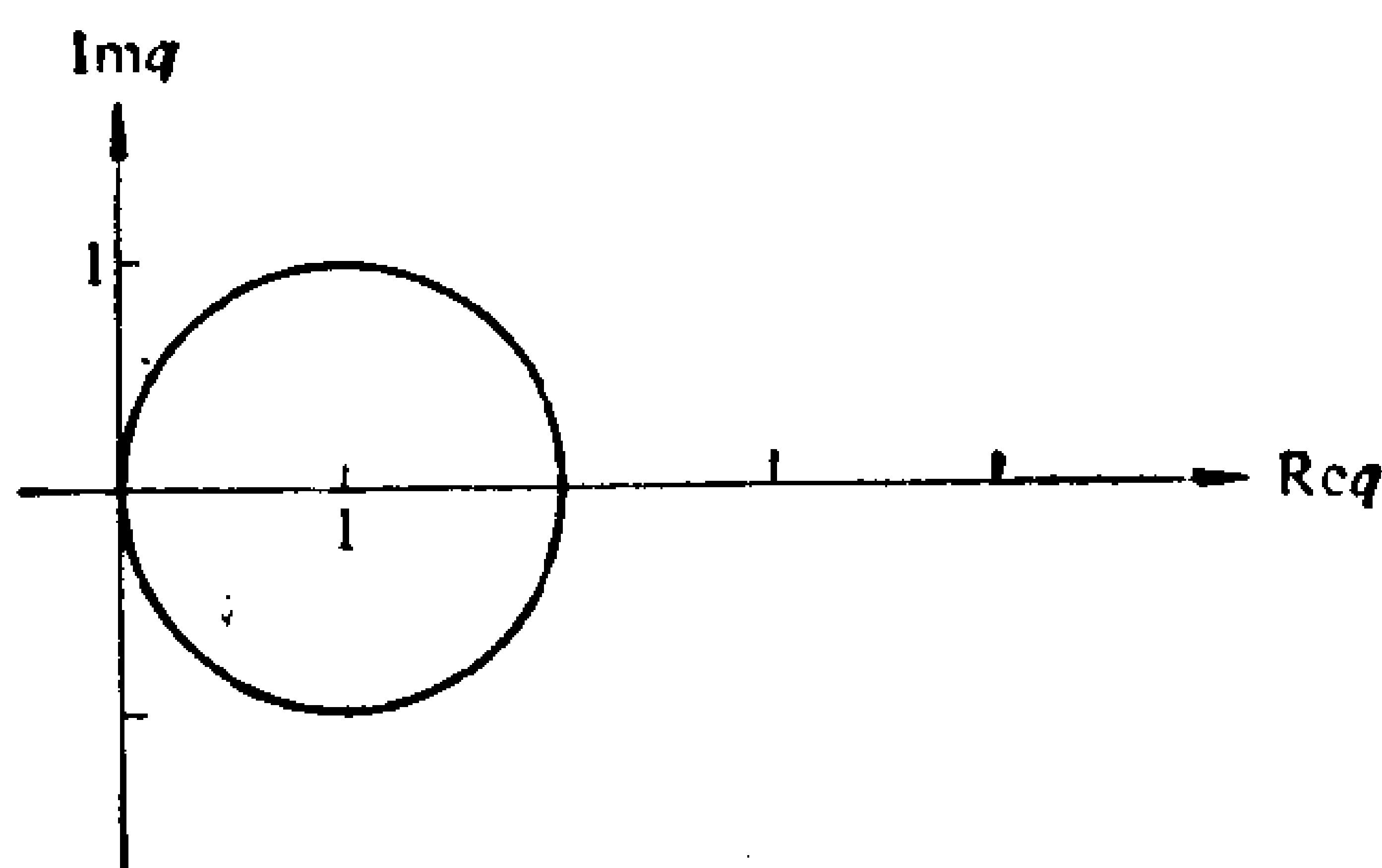


图 3.1 一阶向后差分方法向后 Euler 方法的稳定区域．在闭曲线围成的图形外是稳定的

可以画出使 (3.43) 的根满足根条件的稳定区域的边界. Söderlind[105] 精确地描绘了 $k=1,2,\cdots,8$ 的公式(3.7)的稳定区域的边界,见图 3.1—3.13. 从图中,我们立即看到,当 $k=1,2$ 时,公式(3.7)是 A 稳定的. 当 $k=3,4,5,6$ 时,公式 (3.7) 是刚性稳定的. 当 $k=7,8$ 时,公式(3.7)的多项式 $\rho(\xi)$ 将不满足根条件,因而不是零稳定的.

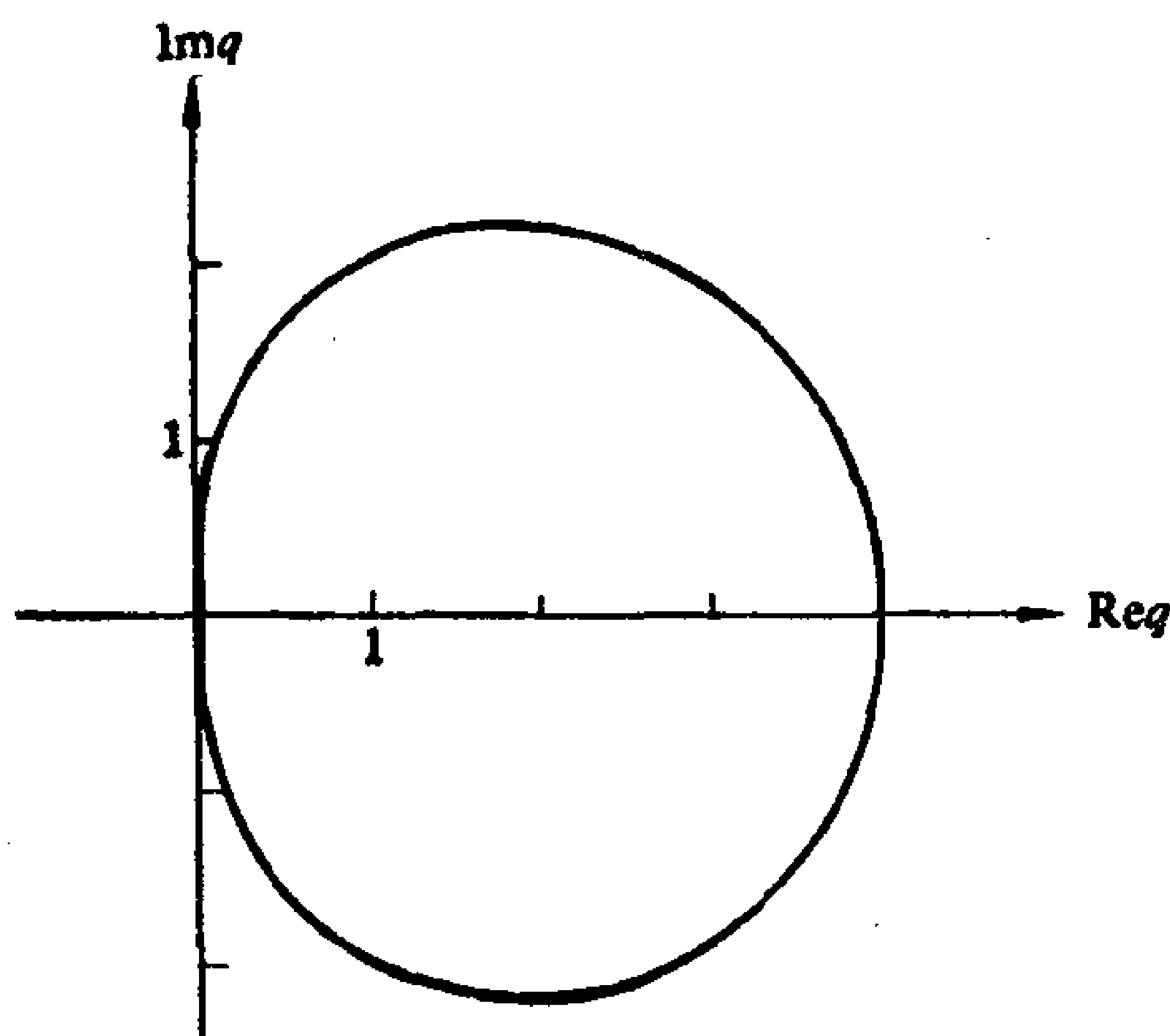


图 3.2 二阶向后差分方法的稳定区域(外部)

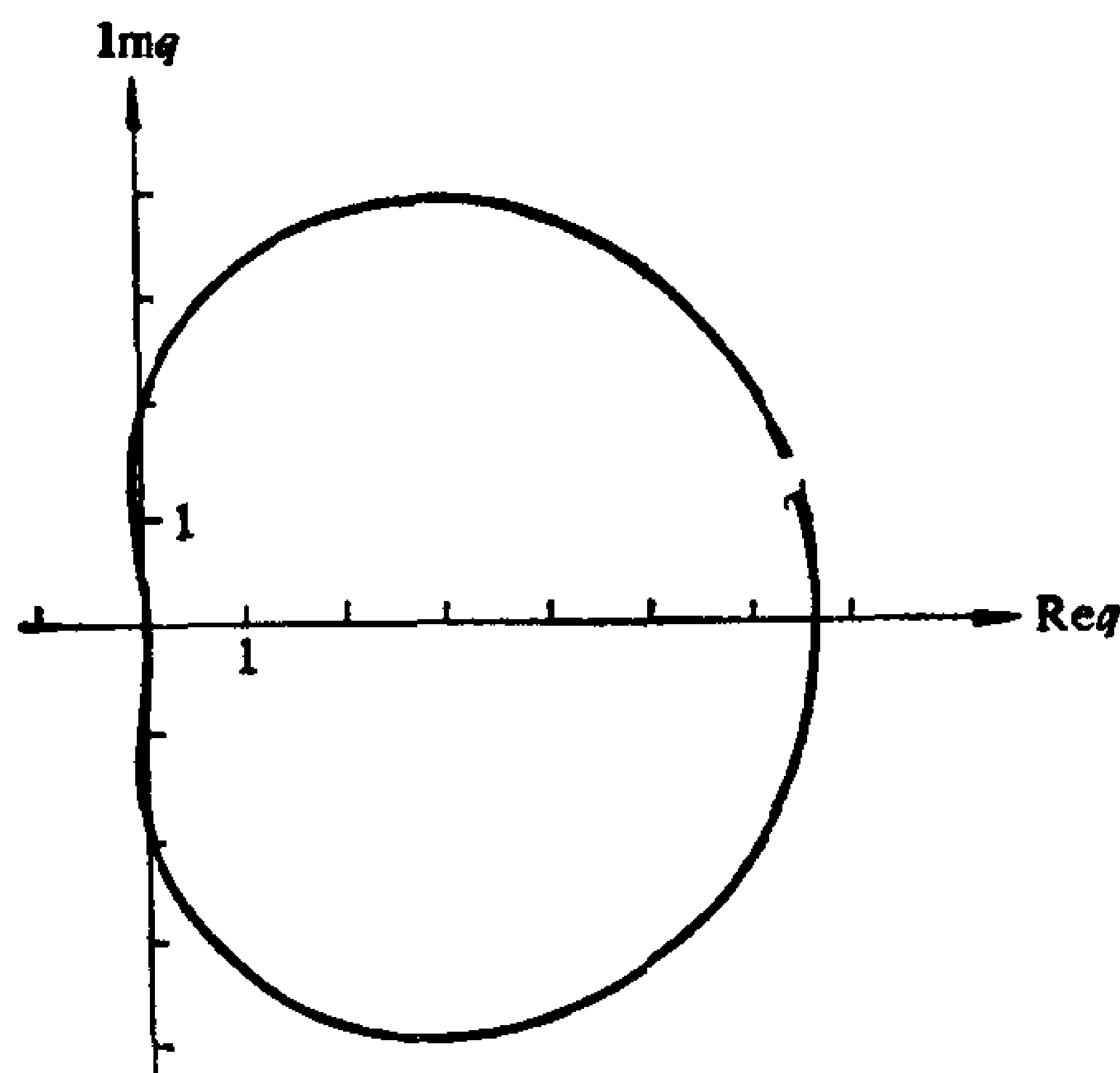


图 3.3 三阶向后差分方法的稳定区域(外部)

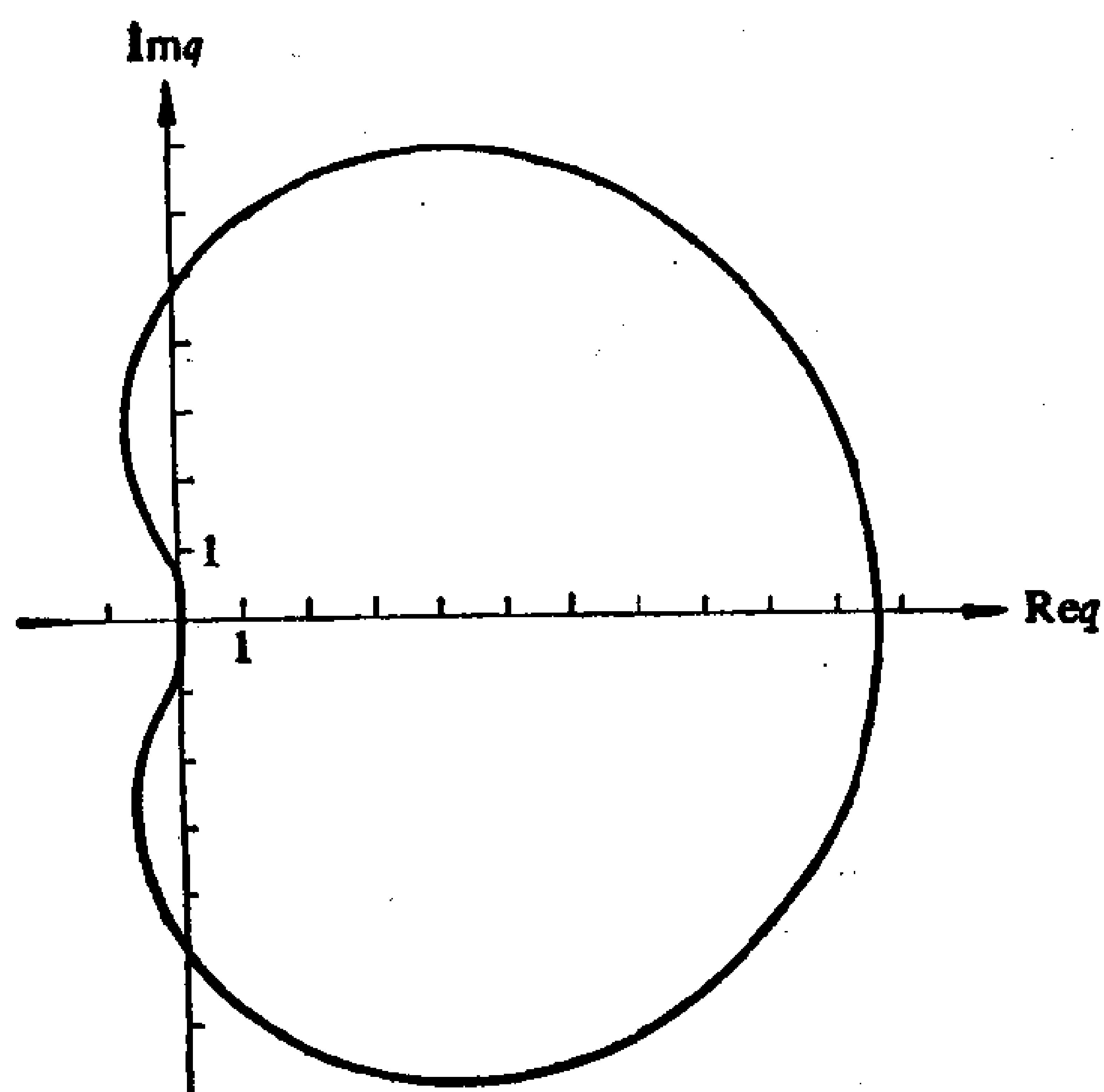


图 3.4 四阶向后差分方法的稳定区域(外部)

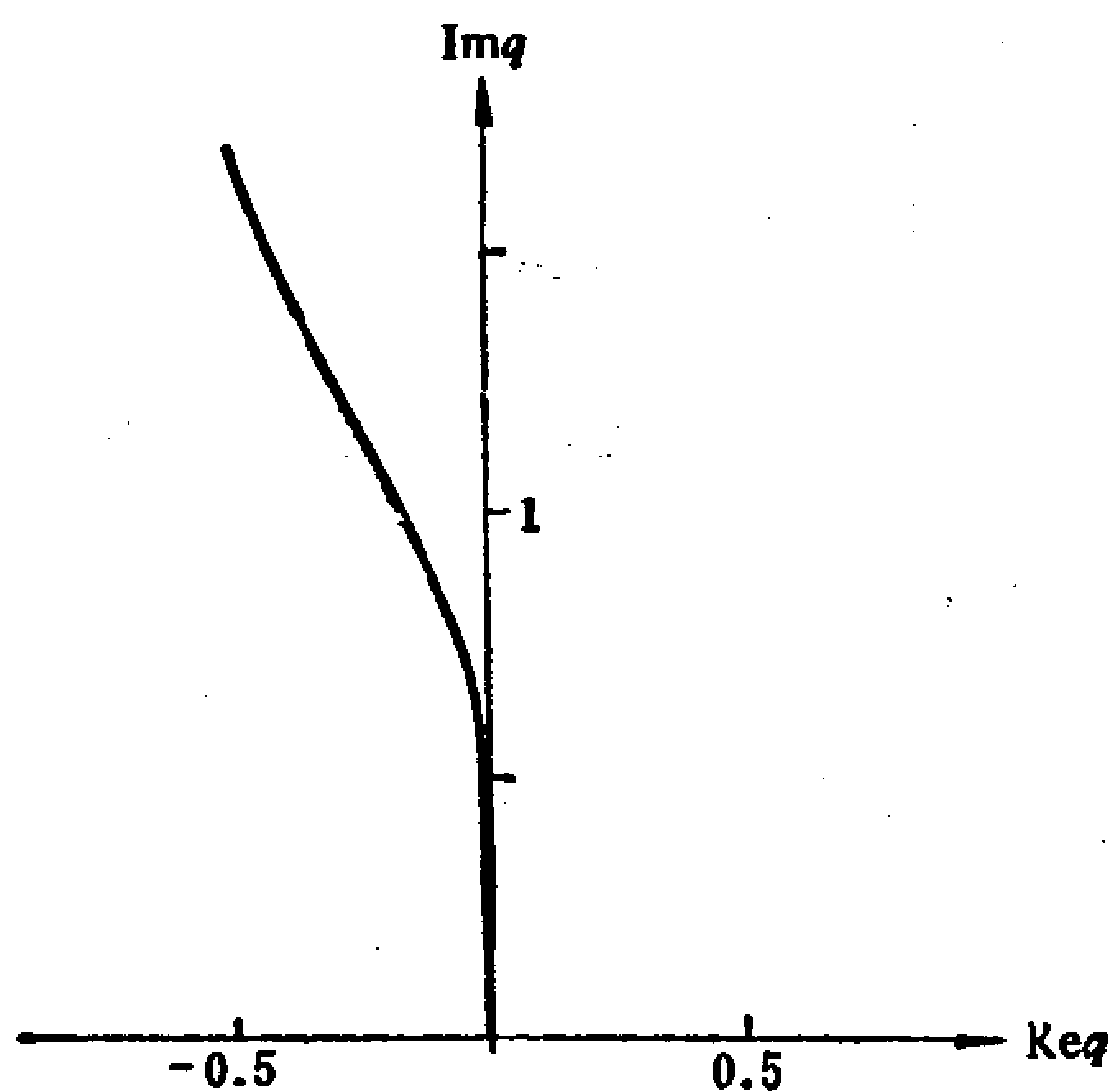


图 3.5 四阶向后差分方法稳定区域在原点附近的边界

Cryer[45] 证明了下面的结果:

定理 3.1 当且仅当 $1 \leqslant k \leqslant 6$ 时,多项式$\rho(\xi)$ 满足根条件.

下面我们来证明这个定理,这里采用 Creedon 和 Miller[44] 给出的证明.

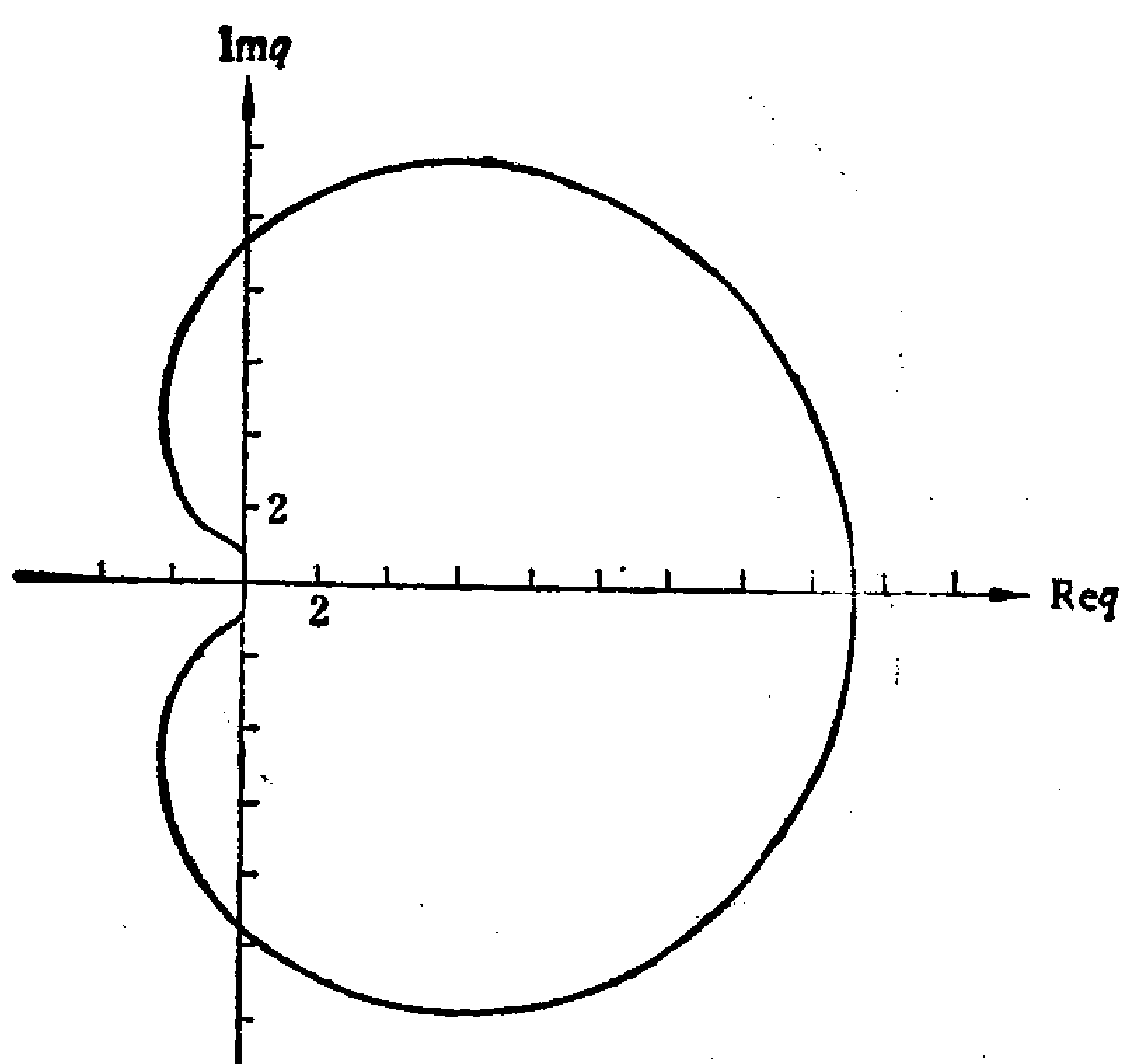


图 3.6 五阶向后差分方法的稳定区域(外部)

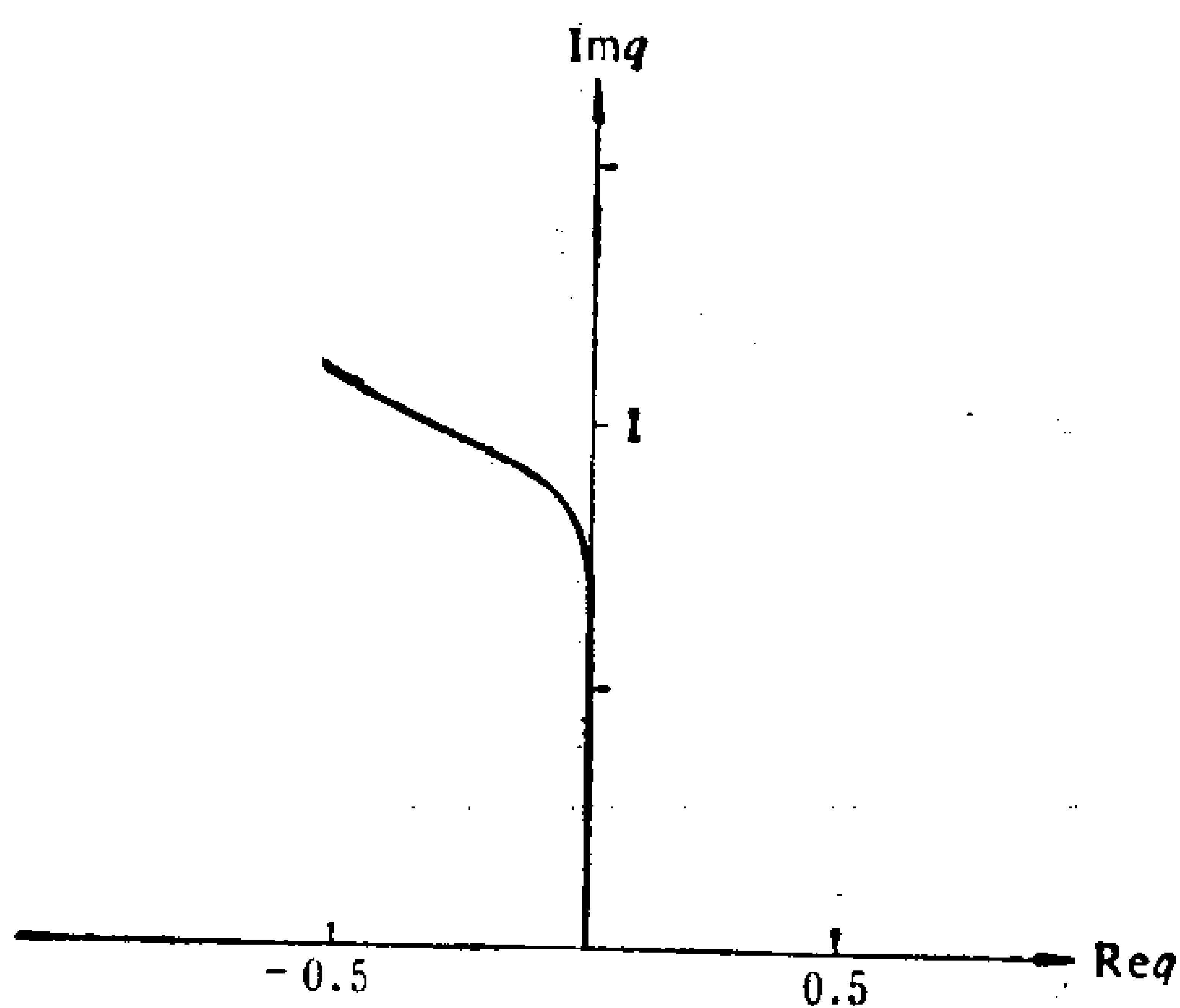


图 3.7 五阶向后差分方法的稳定区域在原点附近的边界

称一个多项式是 (p_1, p_2, p_3) 型的，如果考虑到重数，它有 p_1 个零点在单位圆内，有 p_2 个零点在单位圆上，而有 p_3 个零点在单位圆外. 为了证明定理 1，我们将重复地应用 Miller [82,83, 84, 85] 中给出的判别多项式型的缩减过程.

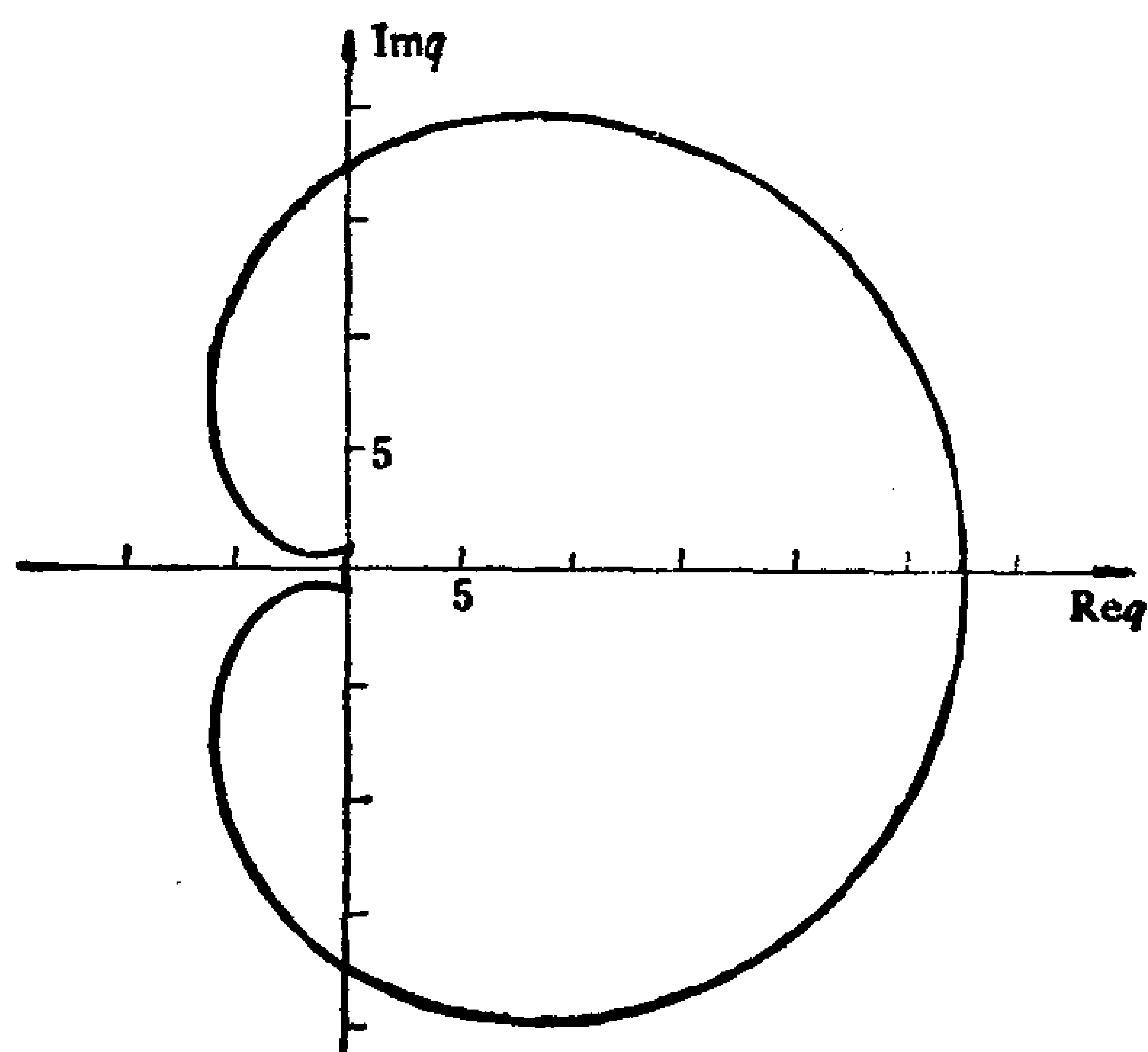


图 3.8 六阶向后差分方法的稳定区域(外部)

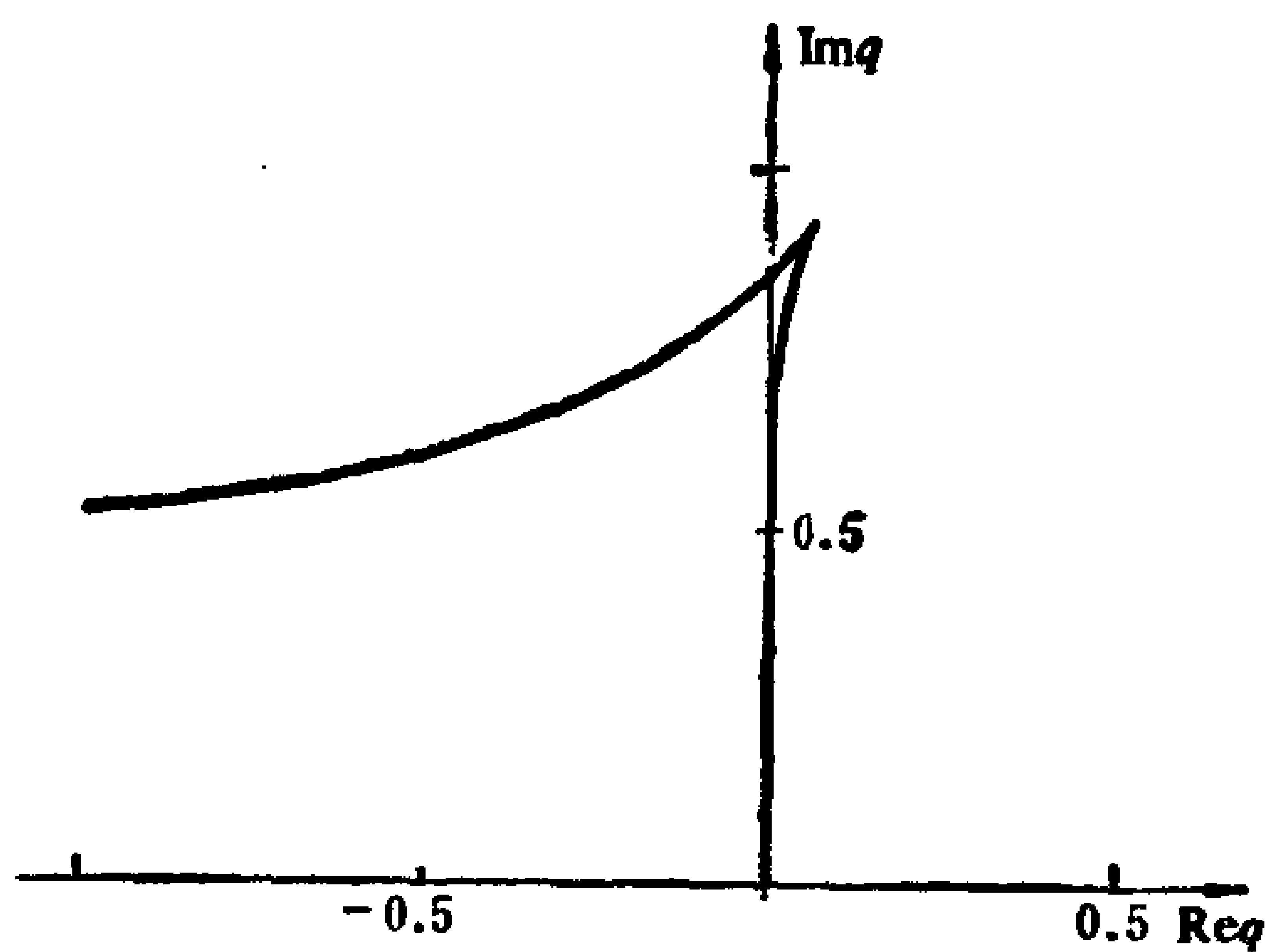


图 3.9 六阶向后差分方法的稳定区域在原点附近的边界

引理 3.2(缩减过程) 令 $\phi(\xi)$ 是具有实(或复)系数的 m 次多项式，令

$$\phi^*(\xi) = \xi^m \overline{\phi(\xi^{-1})}$$

和

$$\hat{\phi}(\xi) = (\phi^*(0)\phi(\xi) - \phi(0)\phi^*(\xi))/\xi,$$

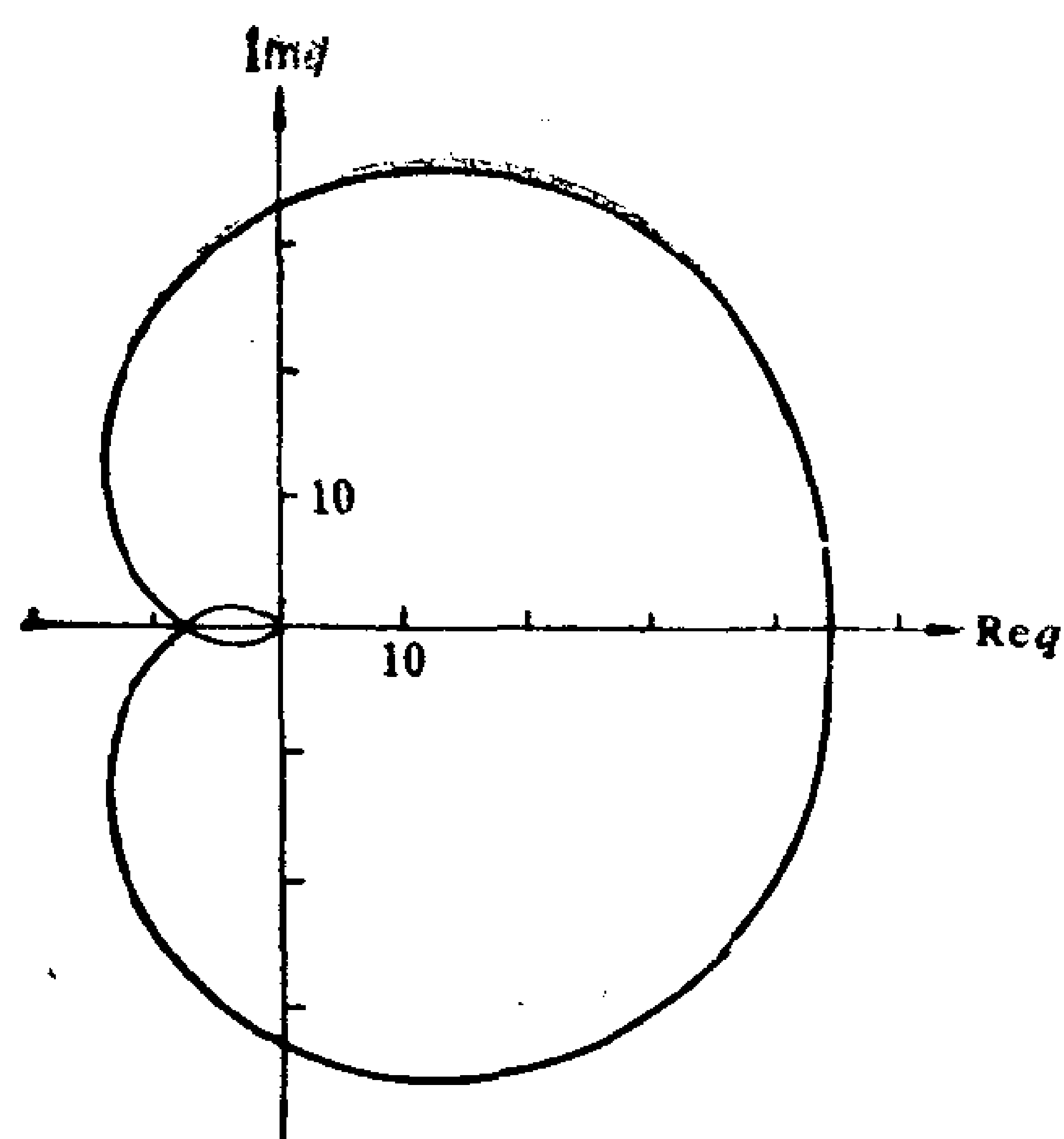


图 3.10　七阶向后差分方法的稳定区域

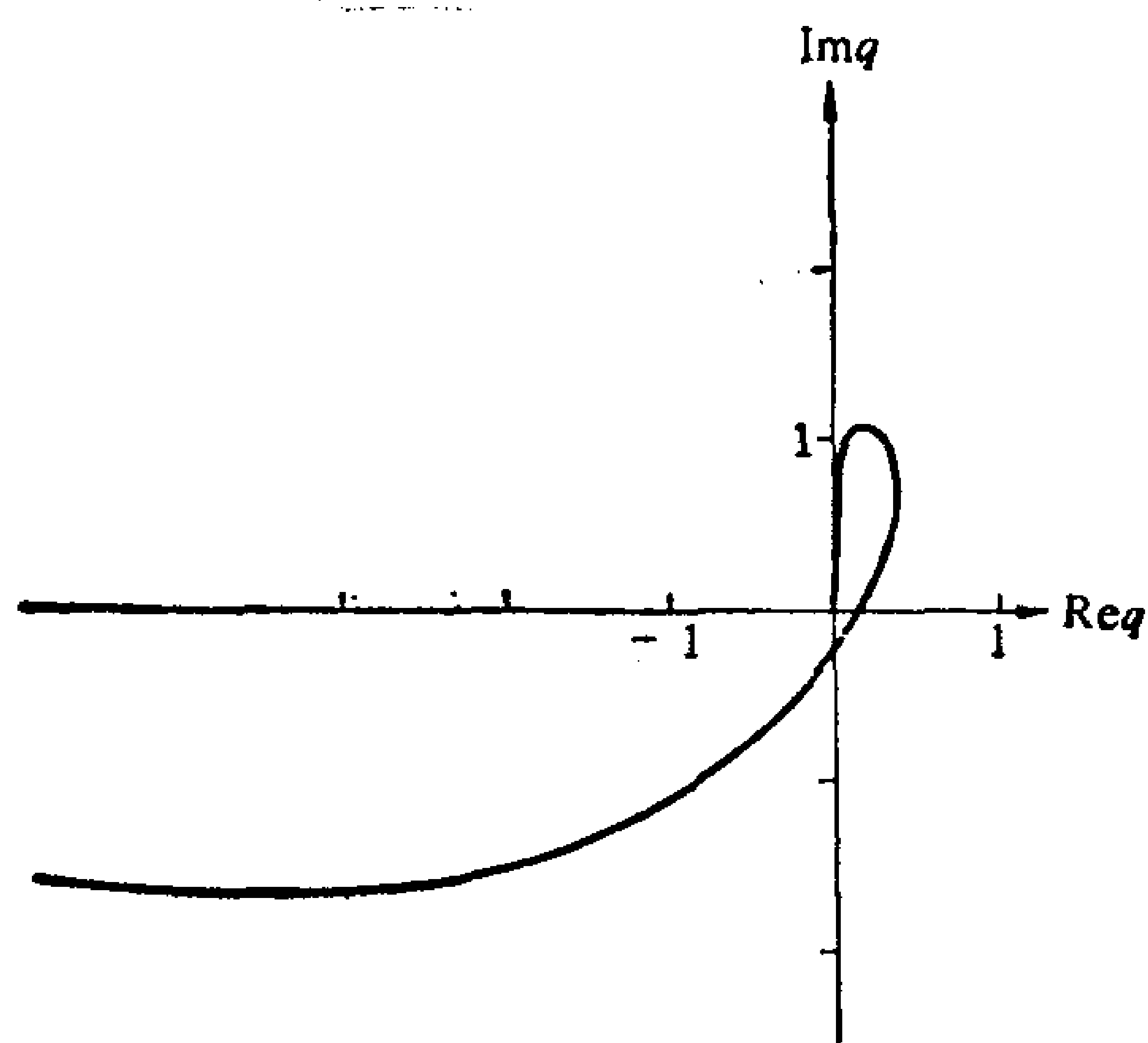


图 3.11　七阶向后差分方法的稳定区域接近原点的边界轨迹

如果$|\phi^*(0)| \neq |\phi(0)|$，则$\phi(\xi)$是型$(p_1, p_2, p_3)$，

$$p_1 + p_2 + p_3 = m$$

的充要条件是或者

(i) $|\phi^*(0)| > |\phi(0)|$和 $\hat{\phi}(\xi)$是$(p_1 - 1, p_2, p_3)$型的，或者

(ii) $|\phi^*(0)| < |\phi(0)|$ 和 $\hat{\phi}(\xi)$ 是$(p_3 - 1, p_2, p_1)$ 型的.

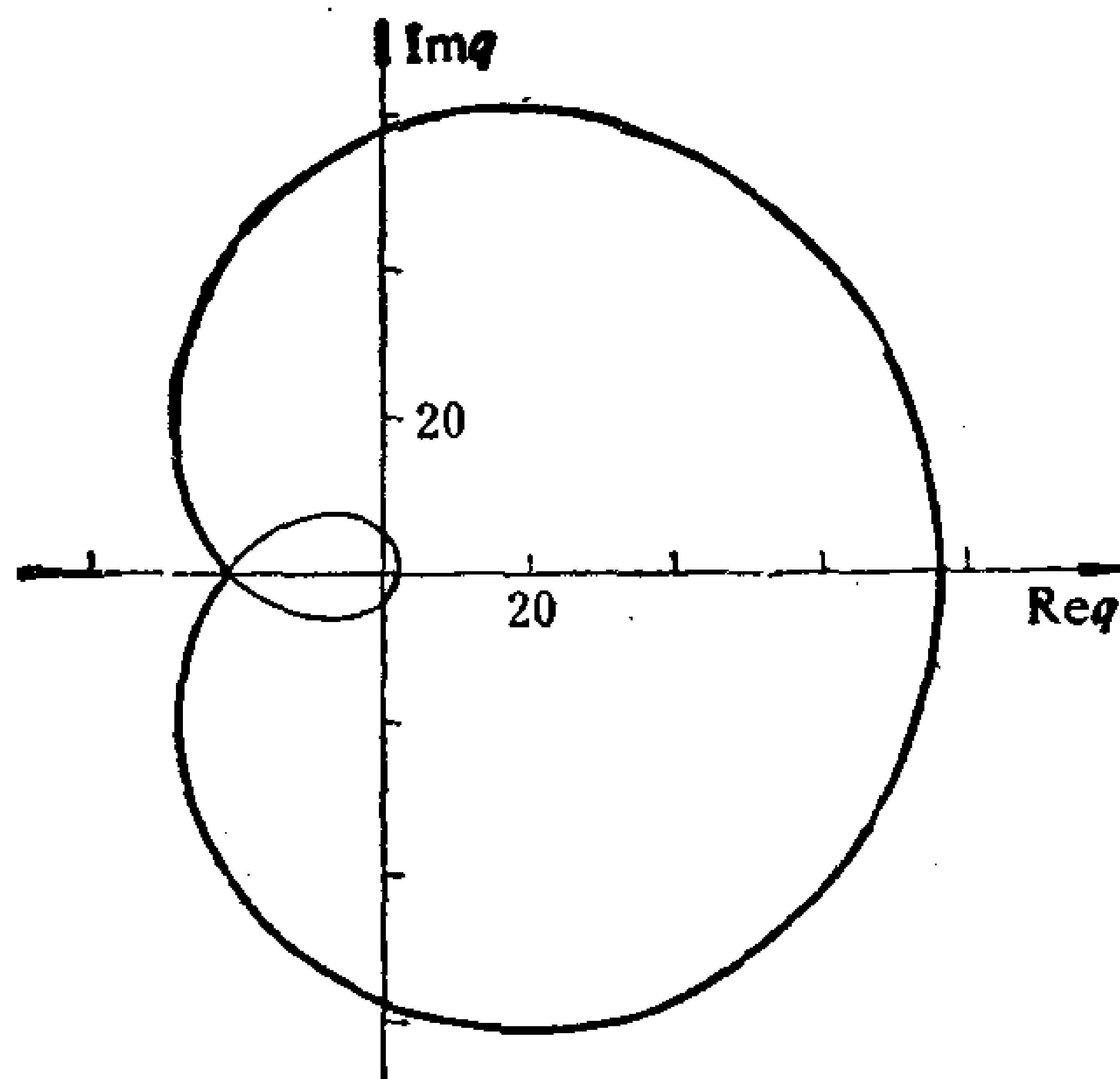


图 3.12 八阶向后差分方法的稳定区域

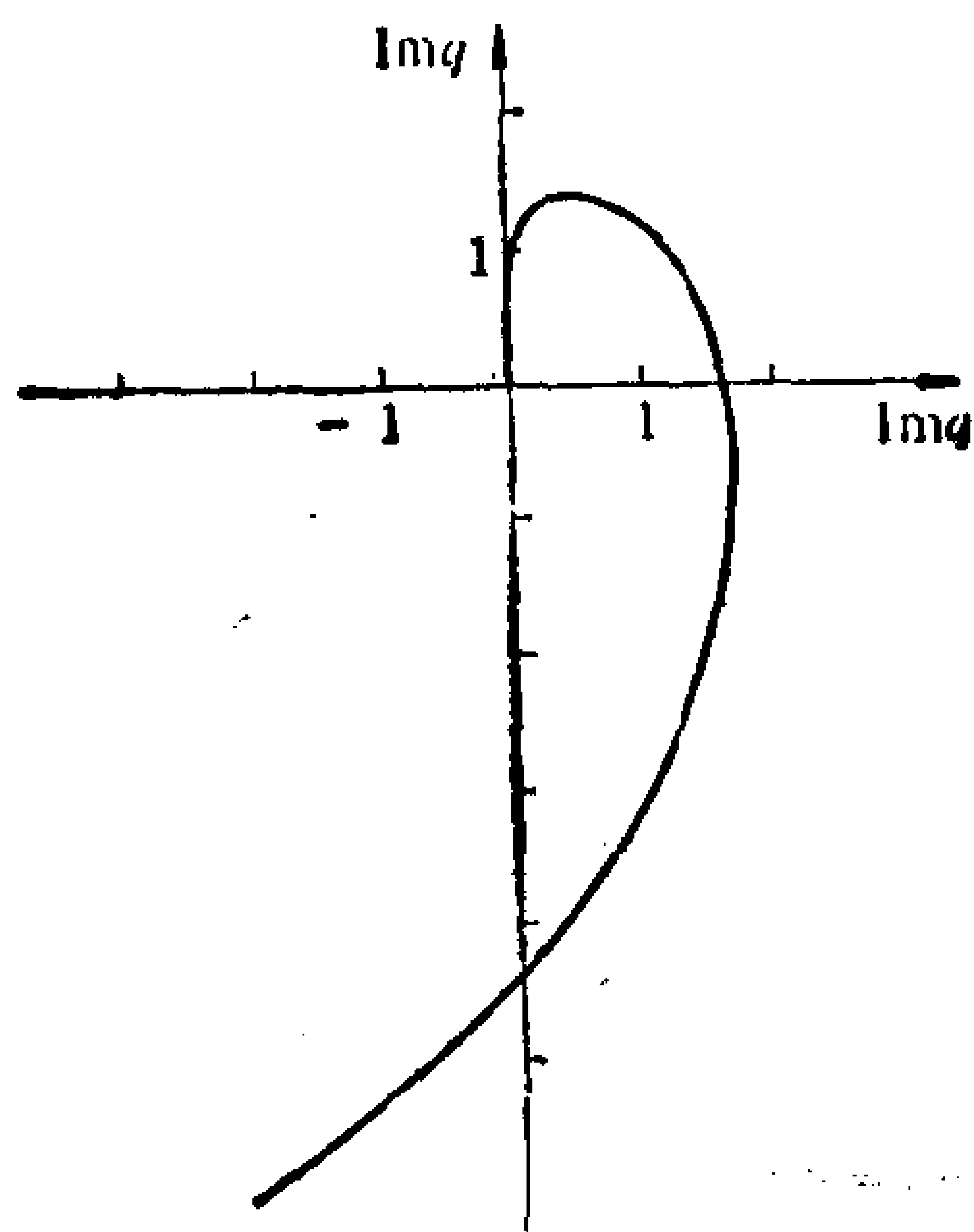


图 3.13 八阶向后差分方法的稳定区域接近于原点的边界轨迹

这个缩减过程的重要点是在所述的假定下，通过考察

$$|\phi^*(0)| - |\phi(0)|$$

的符号，可以将 $\phi(\xi)$ 的型与 $\hat{\phi}(\xi)$ 的型建立联系，并且 $\hat{\phi}(\xi)$ 的次数比 $\phi(\xi)$ 的次数少一.

将定理 3.1 归结成下面两个引理：

引理 3.3 对于 $k \geqslant 13$，多项式 $\rho(\xi)$ 不满足根条件.

引理 3.4 对于 $1 \leqslant k \leqslant 6$，多项式 $\rho(\xi)$ 满足根条件. 对于 $7 \leqslant k \leqslant 12$，多项式 $\rho(\xi)$ 不满足根条件.

下面给出引理 3.3 的证明. 由于每个多项式 ρ 均有根 $\xi = 1$. 引进多项式

$$\phi_{0,k}(\xi) = \rho(\xi)/(\xi - 1),$$

其中$\rho(\xi)$是将(3.40)中的 k 换成 $k+1$ 所得到的多项式. $\phi_{0,k}(\xi)$ 是 k 次的. 对于 $i = 1, \cdots, k$，依次地定义 $k - i$ 次多项式

$$\phi_{i,k}(\xi) = \hat{\phi}_{i-1,k}(\xi).$$

由于 $\rho(\xi) = (\xi - 1)\phi_{0,k}(\xi)$ 和 $\xi = 1$ 不是 $\phi_{0,k}(\xi)$ 的根，可以推出如果 $\phi_{0,k}(\xi)$ 满足根条件，则 $\rho(\xi)$ 也满足根条件. 为了证明引理 3.3，只要证明对所有 $k \geqslant 12$，有

(a) $|\phi_{0,k}^*(0)| > |\phi_{0,k}(0)|$（由缩减过程，这推出如果 $\phi_{1,k}$ 是 $(p_1 - 1, p_2, p_3)$ 型的，则 $\phi_{0,k}(\xi)$ 是 (p_1, p_2, p_3) 型的).

(b) $|\phi_{1,k}^*(0)| > |\phi_{1,k}(0)|$（这推出，如果 $\phi_{2,k}(\xi)$ 是 $(p_1 - 2, p_2, p_3)$ 型的，则 $\phi_{1,k}$ 是 $(p_1 - 1, p_2, p_3)$ 型的).

(c) $|\phi_{2,k}^*(0)| < |\phi_{2,k}(0)|$（这推出，如果 $\phi_{3,k}(\xi)$是 $(p_3 - 1, p_2, p_1 - 2)$ 型的，则 $\phi_{2,k}$ 是 $(p_1 - 2, p_2, p_3)$ 型的. 这时 $p_3 \geqslant 1$，所以至少有一个零点在单位圆外).

为了证明 (a)、(b)、(c)，我们需要 $\phi_{0,k}$ 的前三个和最后三个系数. 记

$$\phi_{0,k}(\xi) = a_0 + a_1\xi + \cdots + a_k\xi^k,$$

并且注意到所有的系数均是 k 的实值函数. 由 $\phi_{0,k}$ 的定义，通过一些代数运算，不难得到

$$a_0 = (-1)^k(k+1)^{-1},$$

$$a_1 = a_0 + (-1)^{k-1}(1 + k^{-1}),$$

$$a_2 = a_0 + (-1)^k(k/2 + (k-1)^{-1} - k^{-1}),$$

和

$$a_k = \sum_{j=0}^{k} (j+1)^{-1} > 0,$$

$$a_{k-1} = a_k - (k+1) < 0,$$

$$a_{k-2} = a_k + (k-4)(k+1)/4 > 0.$$

应用这些关系式，并进行一些运算，我们得到

$$|\phi_{0,k}^*(0)| - |\phi_{0,k}(0)| = a_k - |a_0| > 0.$$

这就证明了 (a)． 类似地，我们有

$$\phi_{1,k}(0) = a_k a_1 - a_0 a_{k-1},$$

$$\phi_{1,k}^*(0) = a_k^2 - a_0^2 > 0 \text{ （由 } (a)),$$

和

$$|\phi_{1,k}^*(0)| - |\phi_{1,k}(0)| = a_k(a_k - 1 - k^{-1}) + 1 - (k+1)^{-2} > 0.$$

这就证明了 (b)．最后，我们得到

$$\phi_{2,k}(0) = (a_k^2 - a_0^2)(a_k a_2 - a_0 a_{k-2}) - (a_k a_1 - a_0 a_{k-1})(a_k a_{k-1} - a_0 a_1),$$

$$\phi_{2,k}^*(0) = (a_k^2 - a_0^2)^2 - (a_k a_1 - a_0 a_{k-1})^2 > 0 \text{ （由 } (b))$$

和

$$\phi_{2,k}^*(0) - (-1)^k \phi_{2,k}(0) = (a_k + |a_0|)\varphi(k),$$

其中

$$\begin{aligned}
\varphi(k) &= (a_k - |a_0|)(a_k^2 - a_0^2 - a_k|a_2| + |a_0|a_{k-2}) \\
&\quad + (|a_{k-1}| - |a_1|)(a_k|a_1| - |a_0||a_{k-1}|) \\
&= h_k^3 - (k/2 + 1 + (k-1)^{-1} + (k+1)^{-1})h_k^2 \\
&\quad + (5k/4 + 3/2 + (2(k-1))^{-1} - k^{-1} \\
&\quad - k^{-2} - (k+1)^{-2})h_k - (k + 1/4 - k^{-1} \\
&\quad - (4(k+1))^{-1} - (k+3)^{-3}),
\end{aligned}$$

$$h_k = \sum_{j=0}^{k} (j+1)^{-1}.$$

如果对于所有的 $k \geqslant 12$，有 $\varphi(k) < 0$，则

$$(-1)^k \phi_{2,k}(0) > \phi_{2,k}^*(0) > 0,$$

(c) 得证．于是我们只须证明

(i) $\varphi(12) < 0$

和

(ii) 对于所有的 $k \geqslant 12$, $\varphi(k+1) < \varphi(k)$.

现在 $\varphi(12) = P(12)$, 其中

$$P(x) = x^3 - (7 + 1/11 + 1/13)x^2 + (16 + 1/2 + 1/22 - 1/12 + 1/12^2 - 1/13^2)x - (12 + 1/4 - 1/12 - 1/52 - 1/13^3).$$

由于 $P(1) < 0$, $P(2) > 0$, $P(3) < 0$ 和 $P(3.2) < 0$, 以及 $h_{12} \in (3, 3.2)$, 推出 $P(h_{12}) < 0$. 这就证明了 (i). 为了证明 (ii), 我们注意 $h_{k+1} = h_k + (k+2)^{-1}$. 因此

$$\varphi(k+1) = h_k^3 - (k/2 + 3/2 + k^{-1} - 2(k+2)^{-1})h_k^2 + (5k/4 + 7/4 - (2k)^{-1} - (k+1)^{-1} - (k+1)^{-2})h_k - (k - (k+1)^{-1} + (k+1)^{-2}).$$

由于对于所有的 $k \geqslant 12$, 有 $h_k \geqslant h_{12} > 3$, 因此推得

$$\varphi(k+1) - \varphi(k) < -(h_k + 1)(h_k - 2)/4 < 0.$$

下面给出引理 3.4 的证明. 由于当 $k = 1$ 时, $\rho(\xi)$ 恰好只有一个零点 $\xi = 1$. 对于 $k = 1$ 的情形, 引理显然成立. 于是只要证明:

(i) 对于 $1 \leqslant k \leqslant 5$, $\phi_{0,k}(\xi)$ 满足根条件.

(ii) 对于 $6 \leqslant k \leqslant 11$, $\phi_{0,k}(\xi)$ 不满足根条件.

通过直接计算多项式 $\phi_{i,k}$, $i = 0, 1, \cdots, k-1$, 重复应用缩减过程,我们来证明 (i) 和 (ii). 对于 $1 \leqslant k \leqslant 5$ 多项式 $\phi_{i,k}$ 在表 3.3 中列出.

为了证明(i),由表 3.3,不难检验,对于满足 $1 \leqslant k \leqslant 5$ 的每一个 k,对于 $i = 0, 1, \cdots, k-1$, 有 $|\phi_{i,k}^*(0)| > |\phi_{i,k}(0)|$. 因此由缩减过程, $\phi_{0,k}$ 是 $(k, 0, 0)$ 型的. 为了证明 (ii), 我们注意到, 首先对于 $k = 6$, 对于 $i = 0, 1, 2, 3$, 有 $|\phi_{i,6}^*(0)| > |\phi_{i,6}(0)|$. 而对于 $i = 4$, 有 $|\phi_{4,6}^*(0)| < |\phi_{4,6}(0)|$. 于是由缩减过程推出 $\phi_{0,6}$ 至少有三个根在单位圆内和至少有一个根在圆外. 其次对于满足 $7 \leqslant k \leqslant 11$ 的每个 k 和 $i = 0, 1, 2$, 有

$$|\phi_{i,k}^*(0)| > |\phi_{i,k}(0)|.$$

而对于 $i=3$，有$|\phi_{3,k}^*(0)| < |\phi_{3,k}(0)|$. 这就推出 $\phi_{0,k}$ 至少有二个根在单位圆内和至少有一个根在单位圆外.

表 3.3

$\phi_{0,1}=(3\xi-1)/2$
$\phi_{0,2}=(11\xi^2-7\xi+2)/6$
$\phi_{1,2}=(117\xi-63)/6^2$
$\phi_{0,3}=(25\xi^3-23\xi^2+13\xi-3)/12$
$\phi_{1,3}=(77\xi^2-67\xi+32)\times 8/12^2$
$\phi_{2,3}=(4905\xi-3015)\times 8^2/12^4$
$\phi_{0,4}=(137\xi^4-163\xi^3+137\xi^2-63\xi+12)/60$
$\phi_{1,4}=(745\xi^3-863\xi^2+685\xi-267)\times 25/60^2$
$\phi_{2,4}=(60467\xi^2-57505\xi+34988)\times 8\times 25^2/60^4$
$\phi_{3,4}=(2432097945\xi-1465169895)\times 8^2\times 25^4/60^8$
$\phi_{0,5}=(147\xi^5-213\xi^4+237\xi^3-163\xi^2+62\xi-10)/60$
$\phi_{1,5}=(21509\xi^4-30691\xi^3+33209\xi^2-21591\xi+6984)/60^2$
$\phi_{2,5}=(2364919\xi^3-2910521\xi^2+2756347\xi-1428885)\times 175/60^4$
$\phi_{3,5}=(20637463\xi^2-17112857\xi+13713664)\times 21509\times 8\times 175^2/60^8$
$\phi_{4,5}=(237840298771473\xi-118485982183743)\times 21509^2\times 8^2\times 175^4/60^{16}$

§3 求解刚性方程的数值方法的计算危险性问题

为了很好地解决求解刚性方程时的计算稳定性问题，在设计数值积分方法时，希望方法的绝对稳定区域尽可能大，以致许多方法的稳定区域包含右半复平面上的一个很大的区域. 但是，在这个区域中，求解的系统本身是不稳定的，从而引起这类方法的"危险性"问题. 所谓"危险性"，是指它把一个不稳定的系统错误地当作稳定系统，因而给出错误的结果.

例如向后 Euler 公式

$$y_{n+1}-y_n=hf(t_{n+1},y_{n+1}), \tag{3.45}$$

其绝对稳定区域是复平面上除去圆 $|h\lambda-1|\leqslant 1$ 以外的整个区

域．将(3.45)应用到试验方程

$$y' = \lambda y, \quad y(0) = 1, \tag{3.46}$$

得

$$y_{n+1} = \frac{1}{1-\lambda h} y_n = \frac{1}{(1-\lambda h)^{n+1}} y_0. \tag{3.47}$$

于是不管 $\mathrm{Re}(\lambda)$ 是正还是负，只要 $|h\lambda - 1| > 1$，则序列 $\{|y_n|\}$ 是递减的．而(3.46)的精确解是

$$y(t) = e^{\lambda t},$$

只有当 $\mathrm{Re}(\lambda) < 0$ 时，它才是递减的．当精确解按指数上升时，即当 $\mathrm{Re}(\lambda) > 0$ 时，数值在某些情况可能是下降的．例如 $h = 1$，$\lambda = 3$ 时，方程(3.46)的精确解是 $y = e^{3t}$，而由公式(3.47)所得的解在 $t = t_n$ 时，$y_n = \frac{1}{(-2)^n}$，如图 3.14 所示．我们本来要得

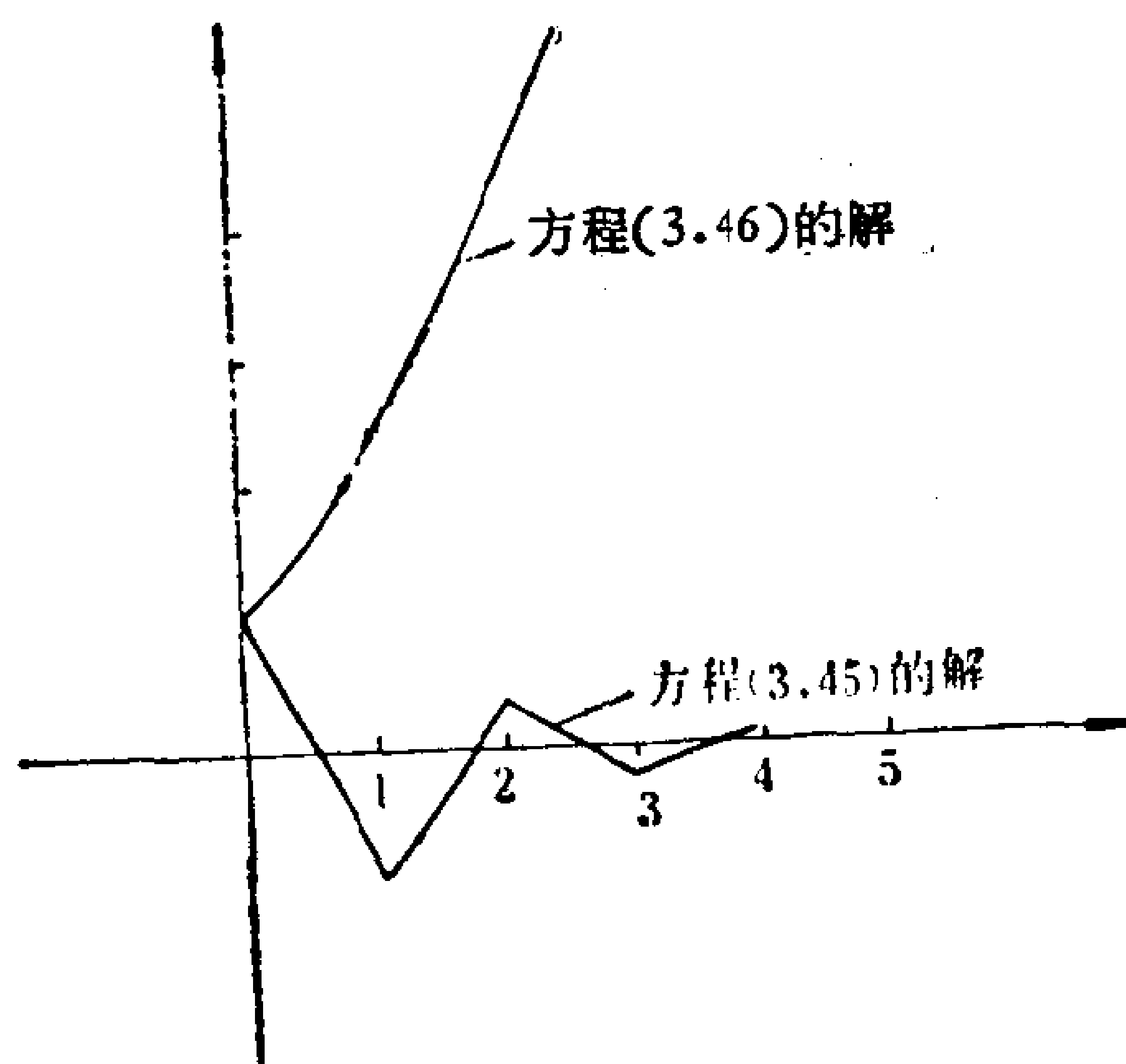


图 3.14　方程(3.46)和方程(3.45)的解

到方程(3.46)的近似解，若方法选择不当，用方程(3.47)所得的结果与所希望求的方程的解毫无关系．这就是所说的计算危险性之一．

由§2，对于其它高阶的向后差分方法大致都有类似的情形．因为此类方法的稳定区域在一个闭曲线的外部，即具有如下的图形

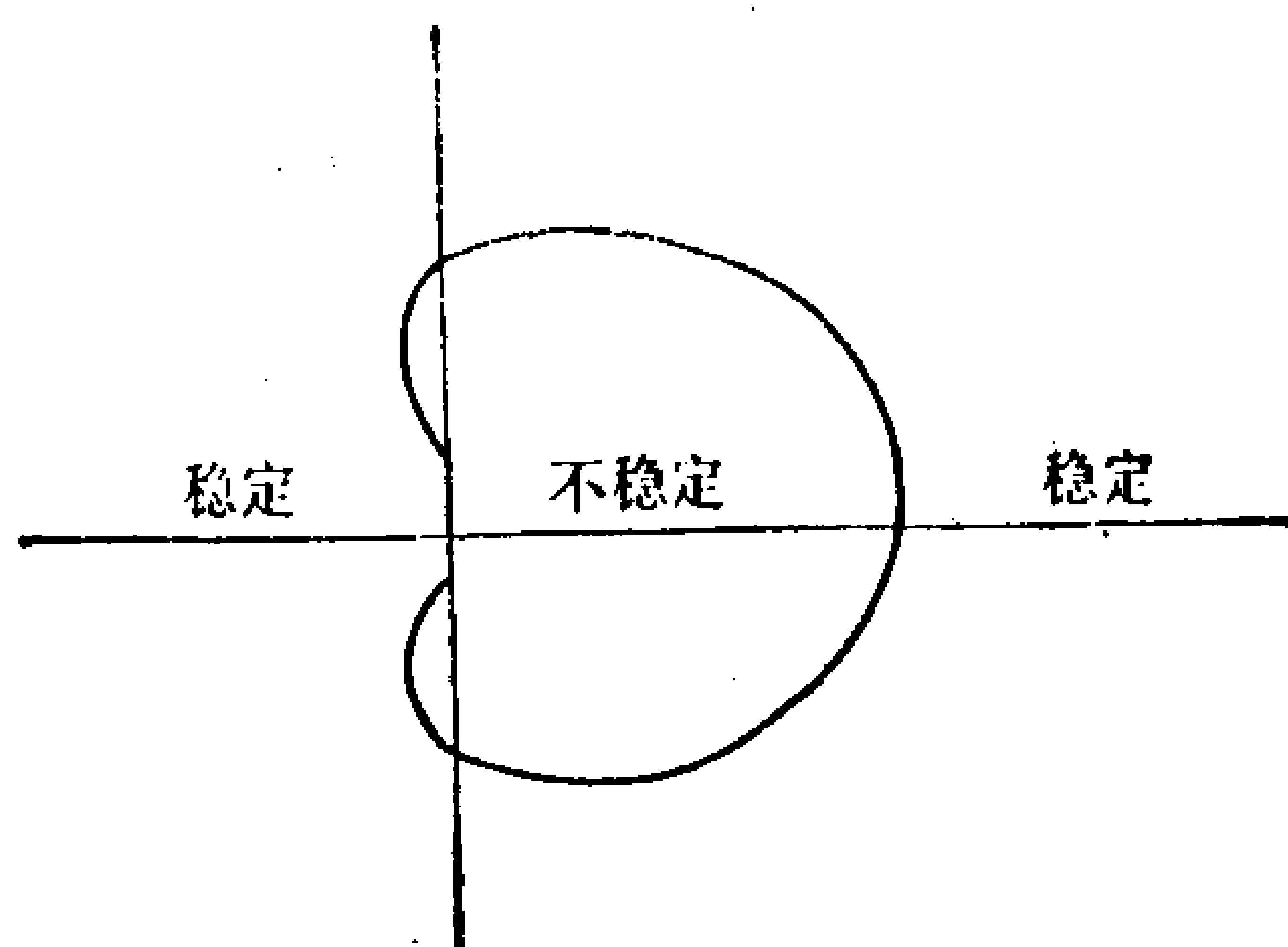


图 3.15

由图看出，方法的稳定区域要包含一部分系统的不稳定的区域。这是产生危险性的最本质的原因。

求解刚性微分方程的程序(例如在§1中描述的 Gear 程序)在每一步计算完成之后，通常都要进行误差检验。根据这种误差估计，自动地选取步长。因此希望这种误差检验过程能发现系统的不稳定性。从而以此作为防止危险性发生的手段。但[76]中构造了试验问题，说明一般的误差检验方法，在某些情况下也发现不了这种不合理的情况。

梯形公式的计算稳定区域和系统的稳定区域是一致的。对于系统(3.46)，其计算公式是

$$y_{n+1}=\left(\frac{1+h\lambda}{1-h\lambda}\right)y_n=\left(\frac{1+h\lambda}{1-h\lambda}\right)^{n+1}y_0.$$

对于不稳定的系统，有

$$\left|\frac{1+h\lambda}{1-h\lambda}\right|>1,$$

因此应该能发现系统的不稳定性。但当 $h\lambda$ 很大时，

$$\left|\frac{1+h\lambda}{1-h\lambda}\right|\sim 1. \tag{3.48}$$

这在计算实践中的反映是，在计算过程中虽然能发现计算公式和系统的不稳定性，但发现的时间太晚，特别是系统的特征值只在某

一个短暂的时间区间上实部变正,可能发现不了. 因此这种情况也是产生危险性的一个源泉.

由上面的讨论我们看到，在构造求解刚性方程的计算方法或计算机程序时,应该考虑被求解的系统是否为稳定的问题. 这个问题到目前还没有得到足够的重视. 下面我们列出[76]中的构造的问题来给出危险性问题的一个直观的说明.

问题 1

$$\begin{aligned}
y_1' &= 10^4 y_1 y_3 + 10^4 y_2 y_4, & y_1(0) &= 1, \\
y_2' &= -10^4 y_1 y_4 + 10^4 y_2 y_3, & y_2(0) &= 1, \\
y_3' &= 1 - y_3, & y_3(0) &= -1, \\
y_4' &= -y_4 - 0.5 y_3 + 0.5, & y_4(0) &= 0, \\
t &\in [0, 10].
\end{aligned}$$

这个问题的精确解可用下面的方式来刻划. 由第 3 第 4 个方程，可推得

$$\begin{aligned}
y_3(t) &= 1 - 2e^{-t}, \\
y_4(t) &= te^{-t}.
\end{aligned}$$

令

$$Y(t) = [y_1(t), y_2(t)]^T,$$

于是上面第 1 和第 2 个方程改写为

$$Y' = A(t)Y, \quad Y(0) = [1,1]^T.$$

其中

$$A(t) = 10^4 \begin{bmatrix} 1 - 2e^{-t} & te^{-t} \\ -te^{-t} & 1 - 2e^{-t} \end{bmatrix},$$

$A(t)$ 的特征值是

$$\lambda_{1,2}(t) = 10^4[(1 - 2e^{-t}) \pm ite^{-t}].$$

在图 3.16 中画出 $\lambda_1(t)$的图形. 对于 Euclid 范数,有

$$\|Y(t)\|_2 = \sqrt{2}\, e^{10^4(2e^{-t}+t-2)}.$$

问题 2

$$\begin{aligned}
y_1' &= -10^4 y_1 y_3 + 10^4 y_2 y_6, & y_1(0) &= 1, \\
y_2' &= -10^4 y_1 y_6 - 10^4 y_2 y_3, & y_2(0) &= 1,
\end{aligned}$$

$$y_3' = -y_3 - y_4 + 1, \qquad y_3(0) = 1.$$
$$y_4' = -2y_4, \qquad y_4(0) = B,$$
$$y_5' = 1 - y_5, \qquad y_5(0) = -1,$$
$$y_6' = -y_0 - 0.5y_5 + 0.5, \quad y_6(0) = 0,$$
$$t \in [0,10].$$

这个问题的精确解可以这样来刻划．由第3—第6个方程得到

$$y_3(t) = 1 + Be^{-2t} - Be^{-t},$$
$$y_4(t) = Be^{-2t},$$
$$y_5(t) = 1 - 2e^{-t},$$
$$y_6(t) = te^{-t},$$

令 $Y(t) = [y_1(t), y_2(t)]^T$，则上面的第1第2个方程改写为

$$Y' = A(t)Y, \quad Y(0) = [1,1]^T,$$

其中

$$A(t) = -10^4\begin{bmatrix} 1 + Be^{-2t} - Be^{-t} & -te^{-t} \\ te^{-t} & 1 + Be^{-2t} - Be^{-t} \end{bmatrix}.$$

$A(t)$ 的特征值是

$$\lambda_{1,2}(t) = -10^4[1 + Be^{-2t} - Be^{-t} \pm ite^{-t}].$$

图3.16中画出 $B = 5$ 的 $\lambda_2(t)$．对于这个问题有

$$\|Y(t)\|_2 = \sqrt{2} \cdot e^{10^4[\frac{1}{2}B(e^{-2t}+1-2e^{-t})-t]}.$$

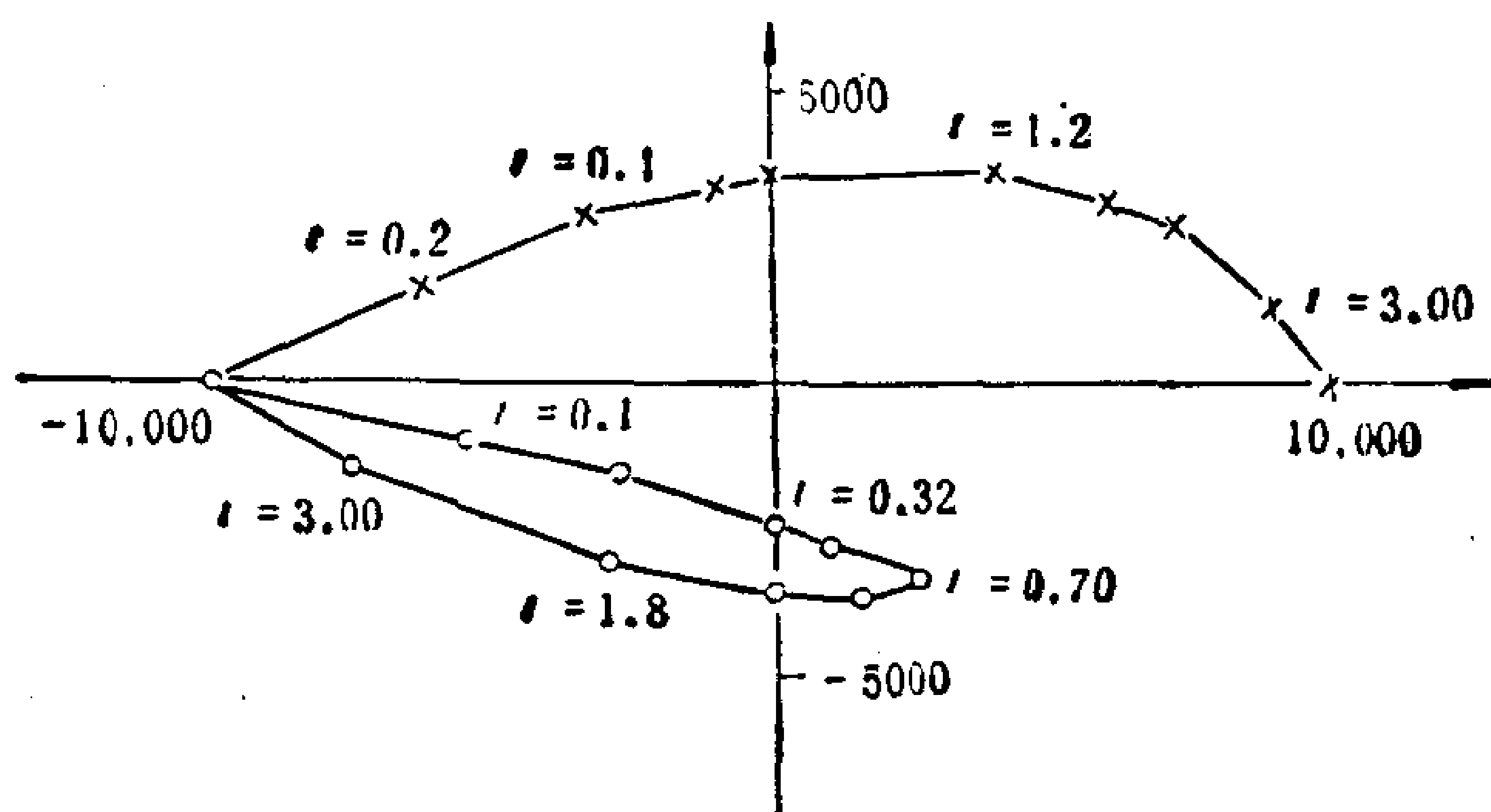


图 3.16　问题1的 $\lambda_1(t)$：——×——×——，
问题2的 $\lambda_2(t)$：——○——○——.

[76]中给出了用 Gear 程序求解问题 1 和问题 2 的结果．我们在表 3.4 和表 3.5 中列出其中的一些结果．$\|Y(t)\|_2$ 是指精确解 $Y(t)$ 的 Euclid 范数，而 $\|\tilde{Y}(t)\|_2$ 是指用 Gear 程序所得的数值解 $\tilde{Y}(t)$ 的 Euclid 范数，$h(t)$ 是所用的步长，控制局部误差的常数是 10^{-6}.

表 3.4　问题 1 的数值结果

t	$\|Y(t)\|_2$	$h(t)$	$\|\tilde{Y}(t)\|_2$
0	1.4	1.3×10^{-9}	1.4
0.10	7.5×10^{-393}	1.4×10^{-2}	5×10^{-10}
0.30	6.5×10^{-949}	4.5×10^{-2}	3×10^{-9}
0.70	4.1×10^{-1333}	5.2×10^{-2}	1×10^{-10}
1.00	3.7×10^{-1148}	5×10^{-2}	2×10^{-9}
1.20	8.6×10^{-859}	8×10^{-2}	3×10^{-11}
1.40	2.0×10^{-464}	8×10^{-2}	1×10^{-9}
1.60	4.2×10^{16}	8×10^{-2}	1×10^{-10}
4.00	1.3×10^{8345}	2×10^{-1}	8×10^{-2}
10.00	1.3×10^{34744}	2×10^{-1}	2×10^{-15}

表 3.5　问题 2 的数值结果 ($B=5$)

t	$\|Y(t)\|_2$	$h(t)$	$\|\tilde{Y}(t)\|_2$
0	1.4	1.3×10^{-9}	1.4
0.10	1.5×10^{-336}	1.2×10^{-2}	2×10^{-10}
0.30	4.1×10^{-574}	2.5×10^{-2}	1×10^{-9}
0.70	4.3×10^{-289}	3.6×10^{-2}	3×10^{-12}
1.00	3.6×10^{-5}	5×10^{-2}	6×10^{-11}
1.20	3.9×10^{90}	7×10^{-2}	1×10^{-9}
1.40	6.9×10^{82}	7×10^{-2}	7×10^{-10}
1.60	1.8×10^{-33}	7×10^{-2}	1×10^{-10}
4.00	4.5×10^{-6909}	1.2×10^{-1}	8×10^{-11}
10.00	1.2×10^{-32572}	1.4×10^{-1}	4×10^{-21}

这两个例子说明，当将适合于求解刚性方程的数值方法应用到特征值从大的负值变化到大的正值的方程组时，将得到错误的结果．

对于这两个问题，初始点的暂态都精确地得到了，这时的步长均是很小的．当对应于大的负实部的特征值的解分量与解的其它分量相比可忽略时，步长开始增加．由这以后，步长由对应于小的特征值的解分量的变化来确定．这时对应于大特征值的分量继续减小．当大特征值在复平面上从左半平面变到右半平面时，这些分量非常小．虽然这时精确解是指数上升的，但是只要 $h\lambda$ 落在方法的稳定区域，这些分量将继续衰减，即得到错误的结果．

§4 广义向后差分公式

形式为

$$\sum_{j=0}^{k} \hat{\alpha}_j y_{n+j} = h\hat{\beta}_k f_{n+k} + h\hat{\beta}_{k+1} f_{n+k+1} \tag{3.49}$$

的公式称作广义 k 步向后差分公式，其中 $\hat{\alpha}_k = 1$ 而其它的系数选成使(3.49)的阶为 $k+1$．假定已得到解 $y_n, y_{n+1}, \cdots, y_{n+k-1}$，则通过下面的各步利用公式(3.49)进行计算：

(i) 由通常的 k 步向后差分公式

$$y_{n+k} - h\beta_k f_{n+k} = -\sum_{j=0}^{k-1} \alpha_j y_{n+j} \tag{3.50}$$

计算得 $\bar{y}_{n+k}$．

(ii) 求解方程

$$y_{n+k+1} - h\beta_k f_{n+k+1} = -\alpha_{k-1}\bar{y}_{n+k} - \sum_{j=0}^{k-2} \alpha_j y_{n+j+1}, \tag{3.51}$$

得到 $\bar{y}_{n+k+1}$．

(iii) 计算 $\bar{f}_{n+k+1} = f(t_{n+k+1}, \bar{y}_{n+k+1})$．

(iv) 由(3.49)的变形

$$y_{n+k} - h\hat{\beta}_k f_{n+k} = -\sum_{j=0}^{k-1} \hat{\alpha}_j y_{n+j} + h\hat{\beta}_{k+1}\bar{f}_{n+k+1} \tag{3.52}$$

计算 y_{n+k}．

如上，用一些不同的公式按一定顺序计算来完成数值积分的一步，一般我们把这个过程叫做完整格式. 格式中所需要解的隐式代数方程一般用修改的 Newton 方法来求解，并且迭代到收敛. 于是我们有

定理 3.2 假定

(i) 公式(3.50)是 k 阶的，

(ii) 公式(3.49)是 $k+1$ 阶的，

(iii) 确定 $\bar{y}_{n+k}$ 和 $\bar{y}_{n+k+1}$ 的隐式代数方程的求解过程是精确的.

那么公式(3.52)是 $k+1$ 阶的.

格式 (i)—(iv) 可以看成是隐式的预估校正格式，其中(3.50)是预估(仍是隐式的)，而(3.52)是校正. 但是这里的校正是部分校正 ($\bar{f}_{n+k+1}$ 未校正). 这样可节省大量的右函数计算的次数. 为了精确地求出(3.49)的解 y_{n+k}，可以应用迭代格式

$$\sum_{j=0}^{k} \hat{\alpha}_j y_{n+j}^{(p)} = h\hat{\beta}_k f(t_{n+k}, y_{n+k}^{(p)}) + h\hat{\beta}_{k+1} f(t_{n+k+1}, y_{n+k+1}^{(p-1)}), \quad (3.53)$$

并且迭代到收敛. 这样计算量要大得多.

在表 3.6 中列出 $1 \leqslant k \leqslant 8$ 的广义向后差分公式的系数，并且列出 $A(\alpha)$ 稳定的近似的 α 角，每个公式均是 $k+1$ 阶的，当 $k=1, 2, 3$ 时，公式是 L 稳定的. 在表3.7中列出 $k=7, 8$ 的通常的向后差分公式的系数. 由于这些公式不满足根条件，故是不能用的.

在图 3.17 中描出 $k=1, \cdots, 8$ 的计算格式的绝对稳定区域. 在所画线的左边的区域是稳定的，并且这些区域关于实轴是对称的. 为了描出这些区域，首先将(3.50)应用到试验方程 $y' = \lambda y$，得到

$$(1-\beta_k q)\bar{y}_{n+k} + \sum_{j=0}^{k-1} \alpha_j y_{n+j} = 0, \ q = h\lambda, \quad (3.54)$$

再应用(3.51)计算 $\bar{y}_{n+k+1}$，得

$$(1-\beta_k q)\bar{y}_{n+k+1}+\alpha_{k-1}\bar{y}_{n+k}+\sum_{j=0}^{k-2}\alpha_j y_{n+j+1}=0, \quad (3.55)$$

将(3.54)代入(3.55),经过整理我们可以得到表示式

$$\bar{y}_{n+k+1}=\sum_{j=0}^{k-1}\gamma_j(q)y_{n+j}, \quad (3.56)$$

其中 $\gamma_j(q)$ 由 β_k, $\{\alpha_j\}$ 和 q 唯一确定. 将(3.52)应用到同样的试验方程,我们有

$$\sum_{j=0}^{k}\hat{\alpha}_j y_{n+j}=q\hat{\beta}_k y_{n+k}+q\hat{\beta}_{k+1}\sum_{j=0}^{k-1}\gamma_j(q)y_{n+j}.$$

这个式子可以改写成形式

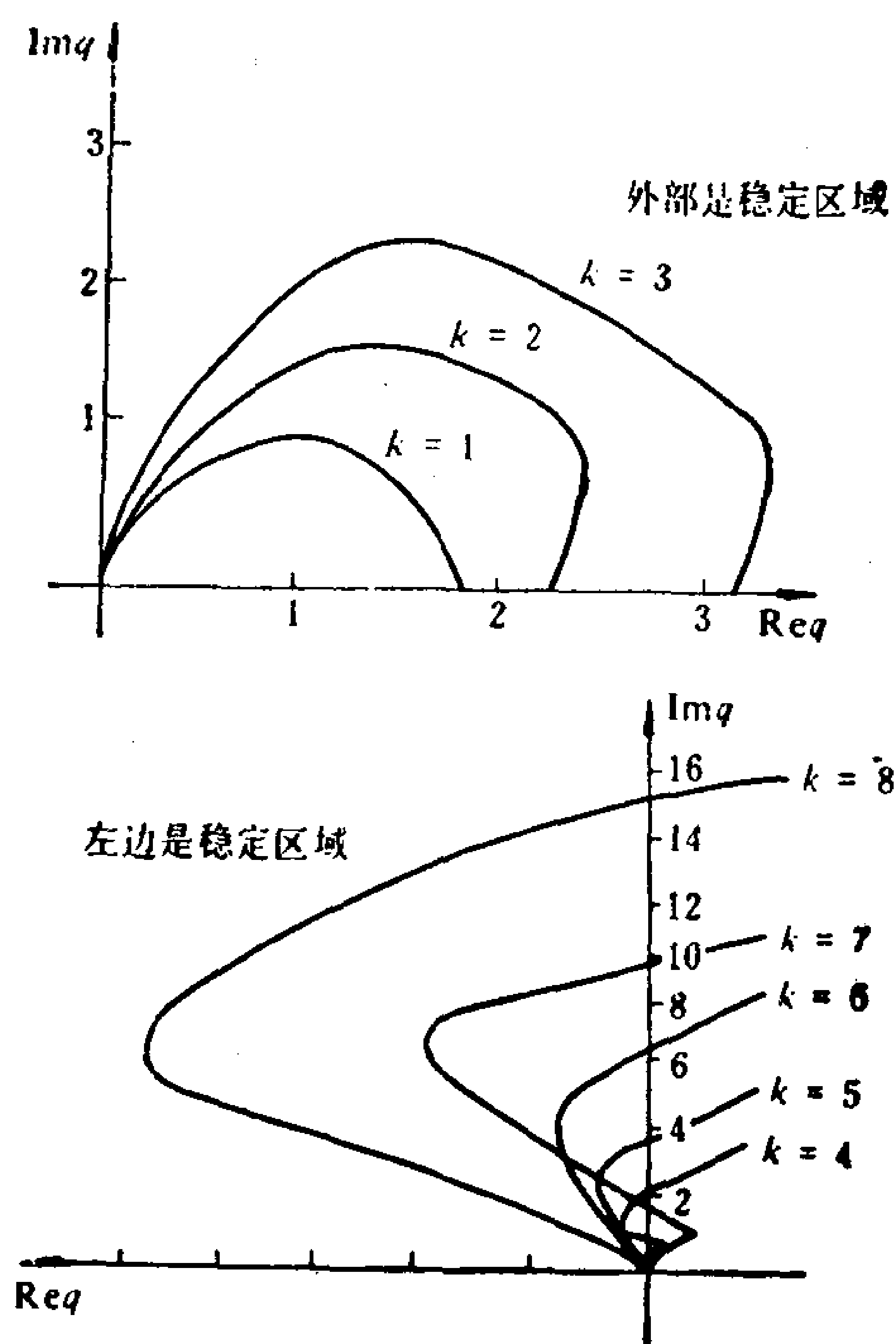


图 3.17 $k=1, 2, \cdots, 8$ 的广义向后差分格式的稳定区域

表 3.6 广义向后差分公式的系数表

k	$\hat{\beta}_{k+1}$	$\hat{\beta}_k$	$\hat{\alpha}_8$	$\hat{\alpha}_7$	$\hat{\alpha}_6$	$\hat{\alpha}_5$	$\hat{\alpha}_4$	$\hat{\alpha}_3$	$\hat{\alpha}_2$	$\hat{\alpha}_1$	$\hat{\alpha}_0$	α
1	$\frac{-1}{2}$	$\frac{3}{2}$								1	-1	90°
2	$\frac{-4}{23}$	$\frac{22}{23}$							1	$\frac{-28}{23}$	$\frac{5}{23}$	90°
3	$\frac{-18}{197}$	$\frac{150}{197}$						1	$\frac{-279}{197}$	$\frac{99}{197}$	$\frac{-17}{197}$	90°
4	$\frac{-144}{2501}$	$\frac{1644}{2501}$					1	$\frac{-4008}{2501}$	$\frac{2124}{2501}$	$\frac{-728}{2501}$	$\frac{111}{2501}$	87°
5	$\frac{-600}{14919}$	$\frac{8820}{14919}$				1	$\frac{-26550}{14919}$	$\frac{18700}{14919}$	$\frac{-9600}{14919}$	$\frac{2925}{14919}$	$\frac{-394}{14919}$	80°
6	$\frac{-1200}{39981}$	$\frac{21780}{39981}$			1	$\frac{-77940}{39981}$	$\frac{68450}{39981}$	$\frac{-46800}{39981}$	$\frac{21375}{39981}$	$\frac{-5756}{39981}$	$\frac{690}{39981}$	67°
7	$\frac{-14700}{626709}$	$\frac{319620}{626709}$		1	$\frac{-1324470}{626709}$	$\frac{1393070}{626709}$	$\frac{-1189475}{626709}$	$\frac{723975}{626709}$	$\frac{-292334}{626709}$	$\frac{70070}{626709}$	$\frac{-7545}{626709}$	50°
8	$\frac{-235200}{12403947}$	$\frac{5988360}{12403947}$	1	$\frac{-28187040}{12403947}$	$\frac{34531280}{12403947}$	$\frac{-35354480}{12403947}$	$\frac{26886300}{12403947}$	$\frac{-14471072}{12403947}$	$\frac{5201840}{12403947}$	$\frac{-1120080}{12403947}$	$\frac{109305}{12403947}$	20°

表 3.7 向后差分公式的系数

k	β_{k+1}	β_k	α_8	α_7	α_6	α_5	α_4	α_3	α_2	α_1	α_0	α
7	—	$\frac{420}{1089}$		1	$\frac{-2940}{1089}$	$\frac{4410}{1089}$	$\frac{-4900}{1089}$	$\frac{3675}{1089}$	$\frac{-1764}{1089}$	$\frac{490}{1089}$	$\frac{-60}{1089}$	
8	—	$\frac{5880}{15981}$	1	$\frac{-47040}{15981}$	$\frac{82320}{15981}$	$\frac{-109760}{15981}$	$\frac{102900}{15981}$	$\frac{-65856}{15981}$	$\frac{27440}{15981}$	$\frac{-6720}{15981}$	$\frac{735}{15981}$	

$$\sum_{i=0}^{k} c_i(q) y_{n+i} = 0,$$

与这个 k 阶差分方程相关的特征方程为

$$\sum_{i=0}^{k} c_i(q) \xi^i = 0.$$

通过它即可画出格式的稳定区域.

由上面我们看出广义向后差分公式比起通常的向后差分公式可以达到更高的精度阶,并且稳定性特征也要好. 这后面一点对于积分 Jacobi 矩阵具有接近于虚轴的模大的特征值的刚性方程是特别重要的. 正是对于这一类问题, Gear 算法的效能是比较差的.

下面我们来考察应用广义向后差分公式的一些处理. 首先我们考察如何有效地求解非线性代数方程, 并且估计每步所需要的计算量. 算法的第一步是计算量 $\bar{y}_{n+k+1}$. 这本质上是应用二次 Gear 方法. 这里确定 $\bar{y}_{n+k}$ 和 $\bar{y}_{n+k+1}$ 的非线性代数方程均是采用拟 Newton 方法求解,一直迭代到收敛. 格式的第二步是应用(3.52)计算 y_{n+k}, 确定 y_{n+k} 的非线性代数方程也用拟 Newton 方法来求解. 实际计算经验表明,由于值 $\bar{y}_{n+k}$ 通常可以作为 y_{n+k} 的非常好的初始近似,这时拟 Newton 方法通常收敛是非常快的. 现在假定解 y_{n+k} 已被接受,将算法往前推进一步计算 y_{n+k+1}. 为此,我们先应用 Gear 方法计算 $\bar{y}_{n+k+1}$ 和 $\bar{y}_{n+k+2}$ (注意 $\bar{y}_{n+k+1}$ 通常可以作为 y_{n+k+1} 的非常好的一次近似). 在得到 $\bar{y}_{n+k+2}$ 以后, 应用 (3.52) 得到 y_{n+k+1}. 按这种方式计算, 我们看到在每一个点 t_{n+k} 计算了 $y(t_{n+k})$ 的三个近似 $\bar{y}_{n+k}$ (由 (3.51)), $\bar{y}_{n+k}$ (由 (3.50)), y_{n+k}. 我们应用的迭代格式分别为

$$(I - h\beta_k J_n)(\bar{y}_{n+k,p} - \bar{y}_{n+k,p-1}) = r_n, \quad 对于 \ \bar{y}_{n+k},$$

$$(I - h\beta_k J_{n+1})(\bar{y}_{n+k+1,p} - \bar{y}_{n+k+1,p-1}) = r_{n+1}, \quad 对于 \ \bar{y}_{n+k+1}, \tag{3.57}$$

和

$$(I - h\hat{\beta}_k\tilde{J}_n)(y_{n+k,p} - y_{n+k,p-1}) = \tilde{r}_n, \text{ 对于 } y_{n+k},$$

其中 r_n, r_{n+1} 和 $\tilde{r}_n$ 均是已知的右边部分，J_n, J_{n+1} 和 $\tilde{J}_n$ 是有关的 Jacobi 矩阵的适当的近似. $y_{n+k,p}$ 是 y_{n+k} 的第 p 次迭代近似,其初始近似按上面描述的得到. 我们注意到，J_n 或 $\frac{\partial f}{\partial y}$ 是在 t_{n+k} 和 $y(t_{n+k})$ 的一个近似上计算 $\frac{\partial f}{\partial y}$ 的值. 而对于 J_{n+1} 是在 t_{n+k+1} 和 $y(t_{n+k+1})$ 的近似处计算的. 可以看出，在每一步中，(3.57) 的第二式所用的 J 可以在下一步中的(3.57)的第一步中应用. 因此,通常在(3.55)中取 $J_n = J_{n+1}$，并且只要收敛就保持这些矩阵不变. 于是,若拟 Newton 迭代格式(3.57)收敛得充分快，完整的算法每一步最多需要计算一个 Jacobi 矩阵和二个系数矩阵的 LU 分解.

第二个要考虑的问题是所用的公式的局部截断误差的估计. 在实际应用时,估计 $\bar{y}_{n+k}$ 中的相对误差,并且将这个误差作为(渐近地)更精确的解 y_{n+k} 中的误差. 这种误差估计经实际计算证明是比较好的. 它类似于 Shampine 和 Gordon 的书[123]中对非刚性情形所用的思想. 因此,我们计算量(按分量)

$$\eta_{n+k} \equiv (y_{n+k} - \bar{y}_{n+k})/\max(1, \bar{y}_{n+k}), \tag{3.58}$$

并将其作为 $\bar{y}_{n+k}$ 的相对误差的估计. 如果这个估计小于预先指定的误差常数 ε，则接受这个点，并求出新的步长 h'，步长 h 和 h' 之间的关系可由下面的程序来确定.

(1) 如果 $\|\eta_{n+k}\| > \varepsilon$，则拒绝 y_{n+k}，并令 $h' = h/2$.

(2) 如果 $\varepsilon > \|\eta_{n+k}\| > \varepsilon/\mu$，其中 $\mu = 2.5 \times 2^{k+1}$，则 $h' = h$，

(3) 如果 $\|\eta_{n+k}\| < \varepsilon/\mu^i, i = 1, 2, 3, 4$，则 $h' = 2^i h$.

[41]对广义向后差分公式(记成 EBD)和通常的向后差分公式(记成 CBD)作了粗略的试验比较. 我们将比较的结果列在下面,试验的条件见[41].

进行试验的问题为

问题 $P1$

$$y_1' = -0.04y_1 + 10000y_2y_3, \qquad y_1(0) = 1,$$
$$y_2' = 0.04y_1 - 10000y_2y_3 - 3\times 10^7 y_2^2, \qquad y_2(0) = 0,$$
$$y_3' = 3\times 10^7 y_2^2, \qquad y_3(0) = 0.$$

$0 \leqslant t \leqslant 40$. 这是 Robertson(1966) 提出的化学中的问题.

问题 $P2$

$$y_1' = -0.013y_1 - 1000y_1y_3, \qquad y_1(0) = 1,$$
$$y_2' = -2500y_2y_3, \qquad y_2(0) = 1,$$
$$y_3' = -0.013y_1 - 1000y_1y_3 - 2500y_2y_3, \qquad y_3(0) = 0,$$

$0 \leqslant t \leqslant 50$. 这是 Gear 1971 年在 *CACM* 中提出的化学中的问题.

问题 $P3$

$$y_1' = 0.01 - (0.01 + y_1 + y_2)(y_1^2 + 1001y_1 + 1001),$$
$$y_1(0) = 0,$$
$$y_2' = 0.01 - (0.01 + y_1 + y_2)(1 + y_2^2), \quad y_2(0) = 0,$$

这是 Liniger 和 Willoughby 提出的反应动力学中的问题(1967).

问题 $P4$

$$y_1' = -\alpha y_1 - \beta y_2 + (\alpha + \beta - 1)e^{-t}, \qquad y_1(0) = 1,$$
$$y_2' = \beta y_1 - \alpha y_2 + (\alpha - \beta - 1)e^{-t}, \qquad y_2(0) = 1.$$

在表 3.8 中列出用相同的步及同样的精度的计算结果. 由表中我们看到 Jacobi 矩阵的计算次数,两种方法几乎是相同的. 而右函数计算的次数 *EBD* 的几乎是 *CBD* 的二倍. 因此,再加上每个 Jacobi 矩阵, *EBD* 需要二次矩阵分解. 所以 *EBD* 的计算量几乎是 *CBD* 的两倍. 但是, *EBD* 达到的整体误差要小得多. 特别是精度常数小的时候更明显. 因而 *EBD* 在精度上超过 *CBD*. 在表 3.9 中列出为达到积分终点同样的整体误差, 两种算法求解问题 $P1$ 所需要的工作量. 从表中可以看到, *EBD* 的工作量比 *CBD* 少得多. 当 $P2$ 整体误差小于 10^{-5} 时, *CBD* 格式所需要的函数计算的次数均大于 10000. 由这个结果看出, 虽然

表 3.8

		CBD			EBD		
	精度	Jacobi 矩阵计算次数	右函数计算次数	整体误差/精度	Jacobi 矩阵计算次数	右函数计算次数	整体误差/精度
问题 $P1$	0.001	19	231	2.9	18	435	0.55
	0.0001	22	474	9.9	24	1035	0.52
	0.00001	28	1173	30.0	32	2725	0.50
问题 $P2$	0.001	13	123	3.5	17	261	1.3
	0.0001	13	243	9.0	14	446	0.008
	0.00001	12	418	55.0	14	871	0.006
	0.000001	10	1308	110.0	11	2729	0.003

表 3.9

得到的整体误差	CBD 所需的工作量		EBD 所需的工作量	
	Jacobi 矩阵的计算次数	右函数的计算次数	Jacobi 矩阵的计算次数	右函数的计算次数
0.00055	26	818	18	435
0.000052	40	3221	24	1035
0.000005		次数超过 10000	32	2725

每步 EBD 的工作量比 CBD 的多，但可用精度高来补偿.

表 3.10 中列出用固定步长 h 求解问题 $P3$ 时积分 10 步的结果. 由表中看出，当步长愈小，EBD 的精度比 CBD 的愈高.

对于问题 $P4$，其 Jacobi 矩阵的特征值是 $-\alpha\pm i\beta$. 而其真解为

$$y_1(t)=y_2(t)=\exp(-t).$$

在表 3.11 中列出了

$$\text{情形 } 1,\ \alpha=\beta=0$$

和

$$\text{情形 } 2,\ \alpha=1,\ \beta=15$$

表 3.10 用固定步长 h 求解问题 P3 积分 10 步的结果

h	真 解	EBD 中的误差	CBD 中的误差
0.01	$y_1 = -0.1096779217 \times 10^{-1}$	0.491×10^{-8}	0.143×10^{-7}
	$y_2 = 0.9879731668 \times 10^{-3}$	0.151×10^{-7}	0.156×10^{-7}
0.001	$y_1 = -0.1006914044 \times 10^{-1}$	0.815×10^{-6}	0.280×10^{-6}
	$y_2 = 0.8978912350 \times 10^{-4}$	0.628×10^{-8}	0.155×10^{-7}
0.0001	$y_1 = -0.6306050198 \times 10^{-2}$	0.835×10^{-6}	0.144×10^{-4}
	$y_2 = 0.3670275606 \times 10^{-5}$	0.819×10^{-6}	0.110×10^{-6}
0.00001	$y_1 = -0.9511426272 \times 10^{-3}$	0.231×10^{-9}	0.268×10^{-7}
	$y_2 = 0.4835591013 \times 10^{-7}$	0.222×10^{-12}	0.261×10^{-10}
0.000001	$y_1 = -0.9949622896 \times 10^{-4}$	0.300×10^{-13}	0.293×10^{-10}
	$y_2 = 0.4983176581 \times 10^{-9}$	0.246×10^{-16}	0.284×10^{-13}

的结果．积分步长为 $h = 0.1$． EBD 用的是三步公式，而 CBD 用的是四步公式．在情形 1，二种结果相仿．而在情形 2，EBD 的结果好得多，这是因为接近于虚轴问题 P4 有大的特征值，并且 CBD 只是阶数不超过 2 时是 L 稳定的，而 EBD 一直到四阶仍是 L 稳定的．

由上面这些比较结果，说明广义向后差分公式似乎是很有希望的一个公式．

§5 应用二阶导数的 Enright 方法

Enright[54] 考虑由常微分方程组

$$\frac{dy}{dt} = f(y), \quad y(0) = y_0 \tag{3.59}$$

给出的自守系统．考虑自守系统的原因是因为实际遇到的许多刚性方程是与时间无关的，即使不是这样，也可以对方程组引进一个附加的方程，将其转化成自守系统． Enright 在通常的 Adams 型线性多步方法中引进二阶导数项，构造了一类 Adams 型 $(k, 2)$ 方

表 3.11

t	CBD 的结果		EBD 的结果		真解
	y_1	y_2	y_1	y_2	
情 形 1					
5.0	0.6721437×10^{-2}	0.6721437×10^{-2}	0.6744973×10^{-2}	0.6744973×10^{-2}	0.6737947×10^{-2}
10.0	0.2873160×10^{-4}	0.2873160×10^{-4}	0.5248640×10^{-4}	0.5248640×10^{-4}	0.4539993×10^{-4}
20.0	-0.1666734×10^{-4}	-0.1666734×10^{-4}	0.7088939×10^{-5}	0.7088939×10^{-5}	0.2061154×10^{-8}
情 形 2					
5.0	0.6542878×10^{-2}	0.6893465×10^{-2}	0.6737930×10^{-2}	0.6737977×10^{-2}	0.6737947×10^{-2}
10.0	0.9031082×10^{-1}	0.6385428×10^{-1}	0.4539983×10^{-4}	0.4540011×10^{-4}	0.4539993×10^{-4}
20.0	-21672.865	-1096.9431	0.2061149×10^{-8}	0.2061162×10^{-8}	0.2061154×10^{-8}

法．其构造算法的出发点是这样的：由 Dahlquist 基本结果，适合于求解刚性方程的线性多步方法一定是隐式的．每积分一步需要求解一个非线性方程组，并且为了避免刚性对步长的限制，通常需要采用 Newton-Raphson 类型的方法．在利用这些方法进行求解时，需要计算(3.59)的右函数的 Jacobi 矩阵 $\partial f(y)/\partial y$． Enright 想，既然在 Newton-Raphson 算法中应用了 Jacobi 矩阵，也可以将这个矩阵放到线性多步公式中去．特别，对于(3.59)，若已经计算了一阶导数 y' 和 Jacobi 矩阵 $\partial f(y)/\partial y$，则立即可以产生二阶导数

$$y'' = (\partial f/\partial y)y', \tag{3.60}$$

因此可以直接将二阶导数构造到方法中去．

Enright[54] 构造了形式为

$$y_{n+k} = \sum_{i=0}^{k-1} \alpha_i y_{n+i} + h \sum_{i=0}^{k} \beta_i y'_{n+i} + h^2 \sum_{i=0}^{k} \gamma_i y''_{n+i} \tag{3.61}$$

的含二阶导数的线性多步公式类．显然，如果 β_k 或 γ_k 是非零的，则公式是隐式的．引进多项式

$$\alpha(\xi) = \xi^k - \sum_{i=0}^{k-1} \alpha_i \xi^i, \tag{3.62}$$

$$\beta(\xi) = \sum_{i=0}^{k} \beta_i \xi^i, \tag{3.63}$$

$$\gamma(\xi) = \sum_{i=0}^{k} \gamma_i \xi^i. \tag{3.64}$$

这时函数 $\Phi(\xi, q)$ 有形式

$$\Phi(\xi, q) = \alpha(\xi) - q\beta(\xi) - q^2\gamma(\xi). \tag{3.65}$$

我们要求选取公式(3.61)中的系数 α_i，β_i，γ_i 使得公式(3.61)是刚性稳定的，并且要满足下面的要求：

(i) 在无穷大处的稳定性．

(ii) 在原点的邻域中有一个合理的稳定性性质（要求刚性稳

定性定义中的 θ 值不太小).

(iii) 具有尽可能高的精度阶.

由 $\Phi(\xi, q)$ 的形式(3.65)可以看出，为了使公式(3.61)在无穷远处是稳定的，$\gamma_k \neq 0$，并且 $\gamma(\xi)$ 的所有根严格地属于单位圆的内部. 因此,为了保证无限远处的稳定性，Enright 取

$$\gamma(\xi) = \gamma_k \xi^k.$$

为了保证在原点的邻域中有一个合理的稳定性性质，与通常的 Adams 公式一样取 $\alpha(\xi) = \xi^k - \xi^{k-1}$. 多项式 $\beta(\xi)$ 的系数 $\beta_i(i = 0, \cdots, k)$ 和 γ_k 选成使公式的阶尽可能的大. 由上述的选取方式推得含二阶导数的线性 k 步公式

$$y_{n+k} = y_{n+k-1} + h\sum_{i=0}^{k} \beta_i y'_{n+i} + h^2 \gamma_k y''_{n+k}, \qquad (3.66)$$

它的阶是 $k+2$，对于 $k \leqslant 7$，系数 β_i 和 γ_k 由表 3.12 给出. 这些公式的稳定区域 R 的边界在图 3.18 上标出. 当 $k=1,2,\cdots,7$ 时公式(3.66)是刚性稳定的. $k=1$ 时公式是 A 稳定的和无限稳定的. 对于 $k=8$ 时稳定区就不是连通的了.

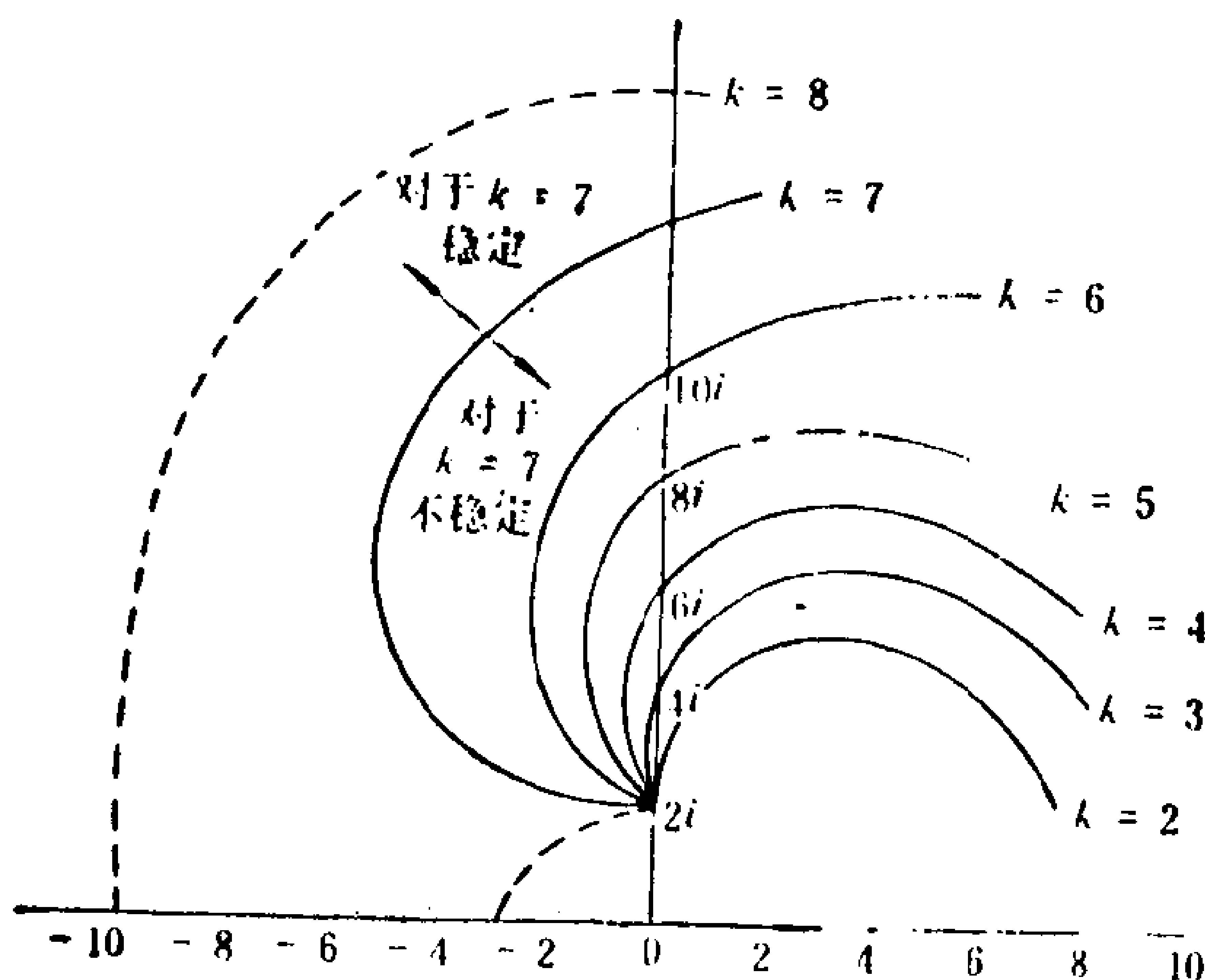


图 3.18 k 步二阶导数多步公式的稳定区域的边界

表 3.12 含二阶导数的 k 步 $k+2$ 阶公式的系数

k	阶 \ 系数	γ_k	β_k	β_{k-1}	β_{k-2}	β_{k-3}	β_{k-4}	β_{k-5}	β_{k-6}	β_{k-7}
1	3	$-\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{3}$						
2	4	$-\frac{1}{8}$	$\frac{29}{48}$	$\frac{5}{12}$	$-\frac{1}{48}$					
3	5	$-\frac{19}{180}$	$\frac{307}{540}$	$\frac{19}{40}$	$-\frac{1}{20}$	$\frac{7}{1080}$				
4	6	$-\frac{3}{32}$	$\frac{3133}{5760}$	$\frac{47}{90}$	$-\frac{41}{480}$	$\frac{1}{45}$	$-\frac{17}{5760}$			
5	7	$-\frac{863}{10080}$	$\frac{317731}{604800}$	$\frac{2837}{5040}$	$-\frac{1271}{10080}$	$\frac{373}{7560}$	$-\frac{529}{40320}$	$\frac{41}{25200}$		
6	8	$-\frac{275}{3456}$	$\frac{247021}{483840}$	$\frac{12079}{20160}$	$-\frac{13823}{80640}$	$\frac{8131}{90720}$	$-\frac{5771}{161280}$	$\frac{179}{20160}$	$-\frac{731}{725760}$	
7	9	$-\frac{33953}{453600}$	$\frac{1758023}{3528000}$	$\frac{1147051}{1814400}$	$-\frac{133643}{604800}$	$\frac{157513}{1088640}$	$-\frac{2797}{36288}$	$\frac{86791}{3024000}$	$-\frac{35453}{5443200}$	$\frac{8563}{12700800}$

为了进一步说明方法的稳定区域，我们在下面的表 3.13 中列出这个方法对应于刚性稳定性定义中的最小的 D 和最大的 θ. 同时为了和前面的 Gear 方法进行比较，在表 3.13 中还列出了 Gear 方法对应于刚性稳定性定义中的最小的 D 和最大的 θ.

公式(3.66)可以实现成变阶变步长的程序. 从三阶的单步法一直可变到九阶的公式. 在实现过程中，估计局部截断误差的策略是很重要的，因为它直接影响方法的可靠性和效率. 由于量 $h\|(\partial f/\partial y)\|$ 可以是十分大的，象通常的预估校正的误差估计的形式可能不行. 可以采用象 Runge-Kntta 方法中常用的先用步长 h 计算 $y(t_{n+k-1}+h)$ 的近似值，再用步长 $\frac{1}{2}h$ 计算二步计算 $y(t_{n+k-1}+h)$ 的近似值，然后由这二个近似值来估计局部误差.

表 3.13 刚性稳定性的 D 和 θ 的对照表

二阶导数公式			阶	Gear 公式		
最小的 D	最大的 θ	k		最小的 D	最大的 θ	k
			1		A稳定	1
			2		A稳定	2
	A稳定	1	3	0.1	0.75	3
	A稳定	2	4	0.7	0.75	4
0.1	2.0	3	5	2.4	0.75	5
0.52	2.0	4	6	6.1	0.5	6
1.4	2.0	5	7	—	—	
2.7	2.0	6	8	—	—	
5.3	1.9	7	9	—	—	

确定计算值 y_{n+k} 是否接受可根据每步的局部误差是否小于预先指定的误差 τ 来决定. 如果局部误差估值的最大分量的量值小于 $\tau(t_{n+k}-t_{n+k-1})$ 则认为 y_{n+k} 可接受，否则拒绝它，重新计算.

步长选取的策略对于一个方法的有效性是很关键的. 因为太小的步长将导致太多的函数求值，而太大的步长将导致拒绝计算值. 因而导致迭代格式中额外的矩阵求逆(或者 LU 分解). 我

们也必须承认步长改变所需要的工作量是很大的，因为它导致矩阵求逆(或 LU 分解)． 因此只有当步长的改变能得到比较大的好处时才进行．鉴于这一点并且考虑到所用的误差估计，我们采取比较保守的策略，即步长的改变只是取 $\frac{1}{2}$ 或者加倍，而且只当在加倍步长的单位步的误差的估计小于 $\frac{\tau}{2}$ 时才采用加倍步长．这种取法一般不会出现拒绝计算值的情形．若出现时，可以取 $\frac{1}{2}$步长，并从三阶公式重新开始． 只当用一个固定的步长计算了 k 步时(k 是所用的公式的步数)，才考虑放大步长．

阶的选取策略主要是根据数值计算的经验．首先根据指定的误差要求选取可能的最高阶．方法的计算开始用三阶公式，并且只当固定步长计算了至少 $k+1$ 步才考虑提高公式的阶．这样的改变直到阶等于限定的最高阶为止．

在实现公式(3.66)时，采用修改的 Newton-Raphson 公式来求解这个非线性代数方程组，在推导时略去含 $\partial^2 f/\partial y^2$ 的项，得到下面的迭代格式

$$
\begin{aligned}
&y_{n+k}^{0} = y_{n+k-1} \\
&y_{n+k}^{l+1} = y_{n+k}^{l} + \Delta_{n+k}^{l+1}, \quad \text{对于 } l = 1, 2, \cdots, 10,
\end{aligned}
\tag{3.67}
$$

其中 Δ_{n+k}^{l+1} 满足

$$
W\Delta_{n+k}^{l+1} = -y_{n+k}^{l} + h\beta_k f(y_{n+k}^{l}) + h^2\gamma_k \frac{\partial f}{\partial y} f(y_{n+k}^{l}) + y_{n+k-1} + h\sum_{i=0}^{k-1}\beta_i y'_{n+i}, \tag{3.68}
$$

而

$$
W \approx \left[I - h\beta_k\left(\frac{\partial f}{\partial y}\right) - h^2\gamma_k\left(\frac{\partial f}{\partial y}\right)^2 \right]. \tag{3.69}
$$

当阶或步长改变时或者当迭代格式收敛得不是足够快时，要重新计算 W．在计算过程中限定最多的迭代次数为 10．若超过 10 次仍不收敛时，需将步长减半，再按新的步长计算．当不等式

$$|\Delta_{n+k}^{i}| < \tau(t_{n+k} - t_{n+k-1})$$

成立时，则认为迭代格式是收敛的．求解(3.68)时，我们可以保存 W^{-1} 或者W的 LU 分解．

Enright 的二阶导数公式(3.66)所对应的$t_{n+i}(i=0,1,\cdots,k)$之间的长度是相等的，即步长是相等的． Sacks-Davis[99] 将公式(3.66)推广成步长是可变化的．二阶导数公式是根据积分

$$y(t_{n+k}) = y(t_{n+k-1}) + \int_{n+k-1}^{t_{n+k}} f(y(t))dt \tag{3.70}$$

来推导的．设对于节点 t_{n+i}，$i=0,1,\cdots,k-1$ 我们有精确解 $y(t_{n+i})$ 的近似 y_{n+i} 和函数 $f(y(t_{n+i}))$ 的近似 $f_{n+i}=f(y_{n+i})$，并令 $h_{n+i}=t_{n+i}-t_{n+i-1}(i=1,\cdots,k)$． 现在要以步长 h_{n+k} 从点 t_{n+k-1} 推进到 t_{n+k}，求出 $y(t_{n+k})$ 的近似 y_{n+k}．作满足

$$\begin{aligned}
&P(t_{n+i}) = f_{n+i},\ i=1,\cdots,k-1,\\
&P(t_{n+k}) = f(y_{n+k}),\\
&P'(t_{n+k}) = f'(y_{n+k})
\end{aligned}$$

的最低次的 Hermite 多项式 $P(t)$． 用 $P(t)$ 代替 (3.70) 中的 $f(y(t))$，则给出含二阶导数的隐式多步公式

$$y_{n+k} = y_{n+k-1} + \int_{t_{n+k-1}}^{t_{n+k}} P(t)dt. \tag{3.71}$$

这个公式通常写成形式

$$y_{n+k} = y_{n+k-1} + h_{n+k}\sum_{i=1}^{k-1}\beta_{n+i}f_{n+i} + h_{n+k}\beta_{n+k}f(y_{n+k}) + h_{n+k}^{2}\gamma_{n+k}f'(y_{n+k}), \tag{3.72}$$

其中系数 β_{n+i}，$i=1,\cdots,k$，γ_{n+k} 均是步长h_{n+i}，$i=1,\cdots,k$ 的零阶的齐次函数．所以 Enright 所用的固定步长的公式是上述的特殊情形．

(3.72)一般是 y_{n+k} 的非线性方程组，必须用某种迭代方法来求解．例如采用 Newton 方法，这时需要 y_{n+k} 的较好的初始值 $y_{n+k,0}$．它可以由公式

$$y_{n+k,0} = y_{n+k-1} + \int_{t_{n+k-1}}^{t_{n+k}} P_0(t)dt \tag{3.73}$$

来计算，其中 $P_0(t)$ 是插值已知点

$$P_0(t_{n+i}) = f_{n+i},\ i = 0, 1, \cdots, k-1,$$

$$P_0'(t_{n+k-1}) = f_{n+k-1}' = \left.\frac{\partial f}{\partial y}\right|_{t=t_{n+k-1}} f_{n+k-1}$$

的最低次的 Hermite 多项式。 有了 $y_{n+k,0}$ 的值，可以不象在(3.67)中那样由 y_{n+k-1} 计算 y_{n+k} 而由 $y_{n+k,0}$ 来计算。令

$$f_{n+k,0} = P_0(t_{n+k}),\ f_{n+k,0}' = P_0'(t_{n+k}).$$

由(3.71)和(3.73)，我们有

$$y_{n+k} = y_{n+k,0} + \int_{t_{n+k-1}}^{t_{n+k}} (P(t) - P_0(t))\,dt\,. \qquad (3.74)$$

现在可以将 $P_0(t)$ 看成是由 $f_{n+1}, f_{n+2}, \cdots, f_{n+k-1}, f_{n+k,0}, f_{n+k,0}'$ 确定的 Hermite 插值多项式。因而 $P(t)$ 与 $P_0(t)$ 的插值点在前面的部分是相同的．可以记

$$\begin{aligned} P(t) - P_0(t) &= \frac{q_{k-1}(t)}{q_{k-1}(t_{n+k})}(f_{n+k} - f_{n+k,0}) \\ &+ \frac{(t - t_{n+k})q_{k-1}(t)}{q_{k-1}(t_{n+k})}\Big[(f_{n+k}' - f_{n+k,0}') \\ &- \frac{q_{k-1}'(t_{n+k})}{q_{k-1}(t_{n+k})}(f_{n+k} - f_{n+k,0})\Big], \end{aligned} \qquad (3.75)$$

其中

$$q_0(t) = 1,$$

$$q_i(t) = \prod_{j=1}^{i} (t - t_{n+k-j}),\ i \geqslant 1.$$

令

$$g_i(j) = \int_{t_{n+k-1}}^{t_{n+k}} (t - t_{n+k})^{j-1} q_i(t)\,dt$$

和

$$d = \frac{q_{k-1}'(t_{n+k})}{q_{k-1}(t_{n+k})} = \sum_{j=1}^{k-1} \frac{1}{t_{n+k} - t_{n+k-j}}.$$

于是由(3.74)和(3.75)，有

$$
\begin{aligned}
y_{n+k} = y_{n+k,0} &+ h_{n+k}\beta_{n+k}(f_{n+k} - f_{n+k,0}) \\
&+ h_{n+k}^2 \gamma_{n+k}(f'_{n+k} - f'_{n+k,0}),
\end{aligned}
\tag{3.76}
$$

其中

$$
\begin{aligned}
h_{n+k}\beta_{n+k} &= \frac{1}{q_{k-1}(t_{n+k})}[g_{k-1}(1) - d g_{k-1}(2)], \\
h_{n+k}^2 \gamma_{n+k} &= \frac{g_{k-1}(2)}{q_{k-1}(t_{n+k})}.
\end{aligned}
\tag{3.77}
$$

对于节点 t_{n+k}，用迭代法求解方程(3.76)的 y_{n+k}，将解从 t_{n+k-1} 推进到 t_{n+k}.

$g_i(j)$ 可以应用三角形数组来确定．从第一行

$$
g_0(j) = -(-h_{n+k})^j/j, \quad j = 1, 2, \cdots, k+2
$$

开始，后一行的元素可以由关系式

$$
g_i(j) = (t_{n+k} - t_{n+k-i})g_{i-1}(j) + g_{i-1}(j+1)
$$

来计算． 每后一行均少一个元素． (3.77) 中只用到前二列的元素，而公式(3.76)的局部截断误差从下面的(3.78)式中可以看到只用到第三列的元素．

(3.76)中的量 $y_{n+k,0}$, $f_{n+k,0}$ 和 $f'_{n+k,0}$ 可以应用插值多项式的各种表示来计算．我们可以得到

$$
\begin{aligned}
y_{n+k,0} = y_{n+k-1} &+ h_{n+k}f_{n+k-1} + \sum_{i=0}^{k-1}(g_i(2) + h_{n+k}g_i(1)) \\
&\times [f_{n+k-1}; f_{n+k-1}; f_{n+k-2}; \cdots f_{n+k-i-1}],
\end{aligned}
$$

$$
\begin{aligned}
f_{n+k,0} = f_{n+k-1} + h_{n+k} \sum_{i=0}^{k-1} q_i(t_{n+k})[&f_{n+k-1}; f_{n+k-1}; \\
&f_{n+k-2}; \cdots; f_{n+k-i-1}]
\end{aligned}
$$

和

$$
\begin{aligned}
f'_{n+k,0} = [f_{n+k-1};\ f_{n+k-1}] &+ \sum_{i=1}^{k-1}\{(h_{n+k}q'_i(t_{n+k}) + q_i(t_{n+k})) \\
&\times [f_{n+k-1}; f_{n+k-1}; f_{n+k-2}; \cdots; f_{n+k-i-1}]\},
\end{aligned}
$$

公式中的记号 [·; ·; ···; ·] 是差商的记号，特别有

$$[f_{n+k-i};f_{n+k-i+1}]=\frac{1}{h_{n+k-i+1}}(f_{n+k-i+1}-f_{n+k-i}),$$

$$[f_{n+k-1};f_{n+k-1}]=f'_{n+k-1},$$

$$[f_{n+k-1};f_{n+k-1};f_{n+k-2}]=\frac{1}{h_{n+k-1}}\times\left[f'_{n+k-1}-\frac{f_{n+k-1}-f_{n+k-2}}{h_{n+k-1}}\right].$$

Sacks-Davis 估计了公式(3.76)的局部截断误差．若令 $y_{n+k-1}(t)$ 为初值问题

$$y'_{n+k-1}(t)=f(y_{n+k-1}(t)),\quad y_{n+k-1}(t_{n+k-1})=y_{n+k-1}$$

的解，则公式(3.76)的局部截断误差为

$$T_{n+k}=\frac{y_{n+k-1}^{(k+2)}(\xi)}{(k+1)!}g_{k-1}(3),\quad \xi\in[t_{n+k-1},t_{n+k}].\tag{3.78}$$

为了将这误差项与 Adams-Moulton 公式的误差项比较，在使用固定步长 h 的情形，(3.78)可以写成为

$$T_{n+k}=\sigma_k h^{k+2}y_{n+k-1}^{(k+2)}(\xi),$$

其中局部误差常数为

$$\sigma_k=\gamma_{k+1}^*-\left(\frac{k}{k+1}\right)\gamma_k^*,$$

γ_k^* 是 Adams-Moulton 方法的系数．表 3.14 中列出了这些常数中最前面的一些值．

表 3.14　二阶导数方法的误差常数

k	0	1	2	3	4	5
γ_k^*	1	$-\frac{1}{2}$	$-\frac{1}{12}$	$-\frac{1}{24}$	$-\frac{19}{720}$	$-\frac{3}{160}$
σ_k	$-\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{72}$	$\frac{7}{1440}$	$\frac{17}{7200}$	$\frac{41}{30240}$

本 章 附 注

§ 1 是根据 Gear 的书[20]编写的.

§ 2 的材料取自 Söderlind [105] 和 Creedon, Miller [85].

§ 3 是根据 Lindberg [76] 和韩天敏,崔可发[6]编写的.

§ 4 的材料取自 Cash [41].

§ 5 的材料主要取自 Enright [54] 和 Sacks-Davis [99].

第四章　e^z 的有理分式近似

许多常微分方程数值积分方法的研究经常可以将其归化成研究指数函数 e^z 的某种有理分式近似．例如将梯形法用于试验方程得

$$y_{n+1} = \frac{1 + \frac{\lambda h}{2}}{1 - \frac{\lambda h}{2}} y_n,$$

而 $\frac{1+z}{1-z} = 1 + z + \frac{1}{2} z^2 + O(z^3) \approx e^z \ (z \to 0)$．反过来，对于每一个 e^z 的有理分式近似，总可以构造与其相应的数值积分方法．由于这种联系，特别是由于数值求解刚性方程对方法研究的推动，使得 e^z 的有理分式近似，而其中的 Padé 近似，成为一个非常活跃的领域．在这一章中，将介绍一些与刚性方程数值方法研究有关的 e^z 的有理分式近似的一些处理方法和结果．

§1　Padé 近似和可接受性

给定在含点 $z = 0$ 的区域中解析的函数 $f(z)$．我们要用有理函数

$$R_{j,k}(z) = \frac{N_{j,k}(z)}{D_{j,k}(z)} \tag{4.1}$$

来近似这个函数，其中 $N_{j,k}(z)$ 和 $D_{j,k}(z)$ 分别为不超过 k 次和 j 次的多项式

$$N_{j,k}(z) = \sum_{m=0}^{K} a_m z^m,$$
$$D_{j,k}(z) = \sum_{m=0}^{j} b_m z^m. \tag{4.2}$$

定义 4.1 如果有

$$R_{j,k}(z) = f(z) + O(|z|^{s+1}) \quad (z \to 0), \tag{4.3}$$

则称 $R_{j,k}(z)$ 是 $f(z)$ 的 s 阶近似。

假定 $N_{j,k}(z)$ 和 $D_{j,k}(z)$ 无公因子，并且 $b_0 = 1$. 我们选取 $j + k + 1$ 个系数 $a_m, m = 0, 1, \cdots, k$ 和 $b_m, m = 1, 2, \cdots, j$ 使下面二个条件成立:

(i) $f(0) = R_{j,k}(0)$,

(ii) 在 $z = 0$ 处，$f(z)$ 和 $R_{j,k}(z)$ 的前 $j + k$ 次导数相等.

设 $f(z)$ 在 $z = 0$ 处的形式幂级数展开式为

$$f(z) = \sum_{m=0}^{\infty} c_m z^m. \tag{4.4}$$

考虑差

$$f(z) - R_{j,k}(z) = \frac{\left(\sum_{m=0}^{\infty} c_m z^m\right)\left(\sum_{m=0}^{j} b_m z^m\right) - \sum_{m=0}^{k} a_m z^m}{\sum_{m=0}^{j} b_m z^m},$$

可以证明,只要选取 a_m 和 b_m 使上式右边的最低次幂为 z^{k+j+1}, 则条件 (i) 和 (ii) 将都能满足. 因此,我们写出

$$\left(\sum_{m=0}^{\infty} c_m z^m\right)\left(\sum_{m=0}^{j} b_m z^m\right) - \sum_{m=0}^{k} a_m z^m = \sum_{m=k+j+1}^{\infty} d_m z^m. \tag{4.5}$$

(4.5)式右边的前 $k + j$ 次幂为零等价于 a_m 和 b_m 满足方程

$$\begin{aligned} &\sum_{m=0}^{j} c_{j+k-s-m} b_m = 0, \quad s = 0, 1, \cdots, j-1; \\ &a_r = \sum_{m=0}^{r} c_{r-m} b_m, \qquad r = 0, 1, \cdots, k; \end{aligned} \tag{4.6}$$

其中约定

$$\begin{aligned} &c_m = 0, \text{ 如果 } m < 0; \\ &b_m = 0, \text{ 如果 } m > j; \\ &b_0 = 1. \end{aligned}$$

(4.6) 是具有 $j+k+1$ 个未知系数的 $j+k+1$ 阶线性方程组。若它有解，则得到 $f(z)$ 的形式为 (4.1) 的近似。将这样构造的 $R_{j,k}(z)$ 称作函数 $f(z)$ 的第 (j, k) 个 Padé 近似，记成 $P_{j,k}(z)$。按表 4.1 排成的表称作 Padé 表。

表 4.1 Padé 表

$P_{0,0}(z)$	$P_{0,1}(z)$	$P_{0,2}(z)$
$P_{1,0}(z)$	$P_{1,1}(z)$	$P_{1,2}(z)$
$P_{2,0}(z)$	$P_{2,1}(z)$	$P_{2,2}(z)$

由上面 $P_{j,k}(z)$ 的构造，成立关系式

$$P_{j,k}(z) = f(z) + O(|z|^{j+k+1}) \quad (z \to 0), \tag{4.7}$$

即 $P_{j,k}(z)$ 是 $f(z)$ 的 $j+k$ 阶近似。由 (4.7) 还可以推出 Padé 近似是唯一的。

现在来给出 Padé 近似的显式表示。显然，形式幂级数 (4.4) 的部分和 $\sum_{m=0}^{k} c_m z^m$ 是 $f(z)$ 的 $(0, k)$ Padé 近似。定义 Hankel 行列式

$$A_k^{(0)} = 1, \quad k \geqslant 0,$$

$$A_k^{(\nu)} = \det\begin{bmatrix} c_k & c_{k-1} & \cdots c_{k-\nu+1} \\ c_{k+1} & c_k & \cdots c_{k-\nu+2} \\ \vdots & \vdots & \vdots \\ c_{k+\nu-1} & c_{k+\nu-2} \cdots & c_k \end{bmatrix} \quad k \geqslant 0, \ \nu \geqslant 1, \tag{4.8}$$

其中约定 $c_{-m} = 0$, $m = 1, 2\cdots$。这些行列式在研究 Padé 近似时将起重要的作用。特别，如果 $A_k^{(\nu)} \neq 0$，则 (4.2) 中 $N_{j,k}(z)$, $D_{j,k}(z)$ 有显式表达式

$$N_{\nu,k}(z) = \frac{1}{A_k^{(\nu)}} \sum_{m=0}^{k} \det\begin{bmatrix} c_m & c_{m-1} \cdots & c_{m-\nu} \\ c_{k+1} & c_k & \cdots c_{k-\nu+1} \\ \vdots & \vdots & \vdots \\ c_{k+\nu} & c_{k+\nu-1} \cdots & c_k \end{bmatrix} z^m, \tag{4.9}$$

$$D_{\nu,k}(z)=\frac{1}{A_k^{(\nu)}}\det\begin{vmatrix}1 & z & z^2 & \cdots & z^\nu\\ c_{k+1} & c_k & c_{k-1} & \cdots & c_{k-\nu+1}\\ c_{k+2} & c_{k+1} & c_k & \cdots & c_{k-\nu+2}\\ \vdots & \vdots & \vdots & & \vdots\\ c_{k+\nu} & c_{k+\nu-1} & c_{k+\nu-2} & \cdots & c_k\end{vmatrix}. \tag{4.10}$$

对于指数函数 $f(z)=e^z$，其 Padé 近似 $P_{j,k}(z)$ 的分子和分母多项式可以表成为显式

$$N_{j,k}(z)=\sum_{m=0}^{k}\frac{(j+k-m)!k!}{(j+k)!m!(k-m)!}z^m, \tag{4.11}$$

$$D_{j,k}(z)=\sum_{m=0}^{j}\frac{(j+k-m)!j!}{(j+k)!m!(j-m)!}(-z)^m. \tag{4.12}$$

在求解刚性方程的数值积分方法的稳定性分析中，往往要考虑到指数函数的有理近似 (4.1) 的某种“可接受性”性质. 根据 Ehle [51]，引进下面的定义:

定义 4.2 (i) 有理函数 $R_{j,k}(z)$ 称作是 $A(\alpha)$ 可接受的，$\alpha\in\left(0,\frac{\pi}{2}\right)$，如果对于所有 $z\in S_\alpha=\{z\,|\,\mathrm{Re}\,z<0, 0\leqslant|\arg(-z)|<\alpha\}$，有

$$|R_{j,k}(z)|\leqslant 1.$$

(ii) 有理函数 $R_{j,k}(z)$ 称作 $A(0)$ 可接受的，如果存在某个 $\alpha\in\left(0,\frac{\pi}{2}\right)$，使 $R_{j,k}(z)$ 是 $A(\alpha)$ 可接受的.

(iii) 有理函数 $R_{j,k}(z)$ 称作 A 可接受的$\left(\text{或称作 } A\left(\frac{\pi}{2}\right)\text{ 可接受的}\right)$，如果对所有 $\alpha\in\left(0,\frac{\pi}{2}\right)$，$R_{j,k}(z)$ 是 $A(\alpha)$ 可接受的.

定义 4.3 有理函数 $R_{j,k}(z)$ 称作是 $L(\alpha)$ 可接受的，$\alpha\in\left(0,\frac{\pi}{2}\right)$，如果 $R_{j,k}(z)$ 是 $A(\alpha)$ 可接受的，并且有

$$|R_{j,k}(z)|\to 0,\ |z|\to\infty.$$

当我们建立有理函数的可接受性质时，通常应用解析函数的

极大模原理．应用这个原理，得到下面的定理

定理 4.1 给定形式为(4.1)的有理函数 $R_{j,k}(z)$，如果有

(i) $|R_{j,k}(iy)| \leqslant 1,\ y \in R,$ (4.13)

(ii) $\lim\limits_{|z| \to \infty} |R_{j,k}(z)| \leqslant 1,$

(iii) 在集合 $C^- = \{z \in C|, \mathrm{Re} z < 0\}$ 中，$R_{j,k}(z)$ 是解析的．

则 $R_{j,k}(z)$ 是 A 可接受的．

由这个定理，为了验证给定的 $R_{j,k}(z)$ 的 A 可接受性，只要验证定理 4.1 中的条件 (i) (ii) (iii) 是否成立．这三个条件中，条件 (ii) 验证起来比较容易．其它二个条件验证起来比较困难．在 §2 和 §3 中，将提供一些处理验证 (ii) 和 (iii) 的工具和结果

类似于定理 4.1，可以给出相应于 $A(\alpha)$ 可接受的结果．

为研究有理分式的可接受性质，可以改成研究集合

$$A = \{z \in C \mid |R_{j,k}(z)| > |e^z|\} = \{z \in C \mid |S(z)| > 1\} \quad (4.14)$$

的性质，其中

$$S(z) = R_{j,k}(z)/e^z. \quad (4.15)$$

事实上，有下面的定理（在这一节的其余部分若不引起混淆，将省掉 $R_{j,k}(z)$ 的下标。）

定理 4.2 $R(z)$ 是 A 可接受的充分必要条件为

(i) 集合 A 与虚轴不相交，

(ii) $R(z)$ 在 C^- 中无极点．

证明 这个定理由极大模原理及下面的一些事实推得，即 $R(z)$ 与 $S(z)$ 具有相同的零点和极点，在虚轴上 $|e^z| = 1$，和在虚轴上集合 A 和集合

$$D = \{z \in C \mid |R(z)| \leqslant 1\} \quad (4.16)$$

相补．证完．

在图 4.1 中描绘了指数函数 e^z 的 Padé 近似的集合 A 的一些例子．由这些图形立即可以看出哪些是 A 可接受的，而哪些不是 A 接受的．特别可看出，当 $j-2 \leqslant k \leqslant j$ 时 e^z 的 Padé 近似是 A 可接受的(见定理 4.19) 由于我们主要关心指数函数 e^z 的有

理分式近似，在本书的后面部分除特殊指出者外，Padé 近似均是指对 e^z 的.

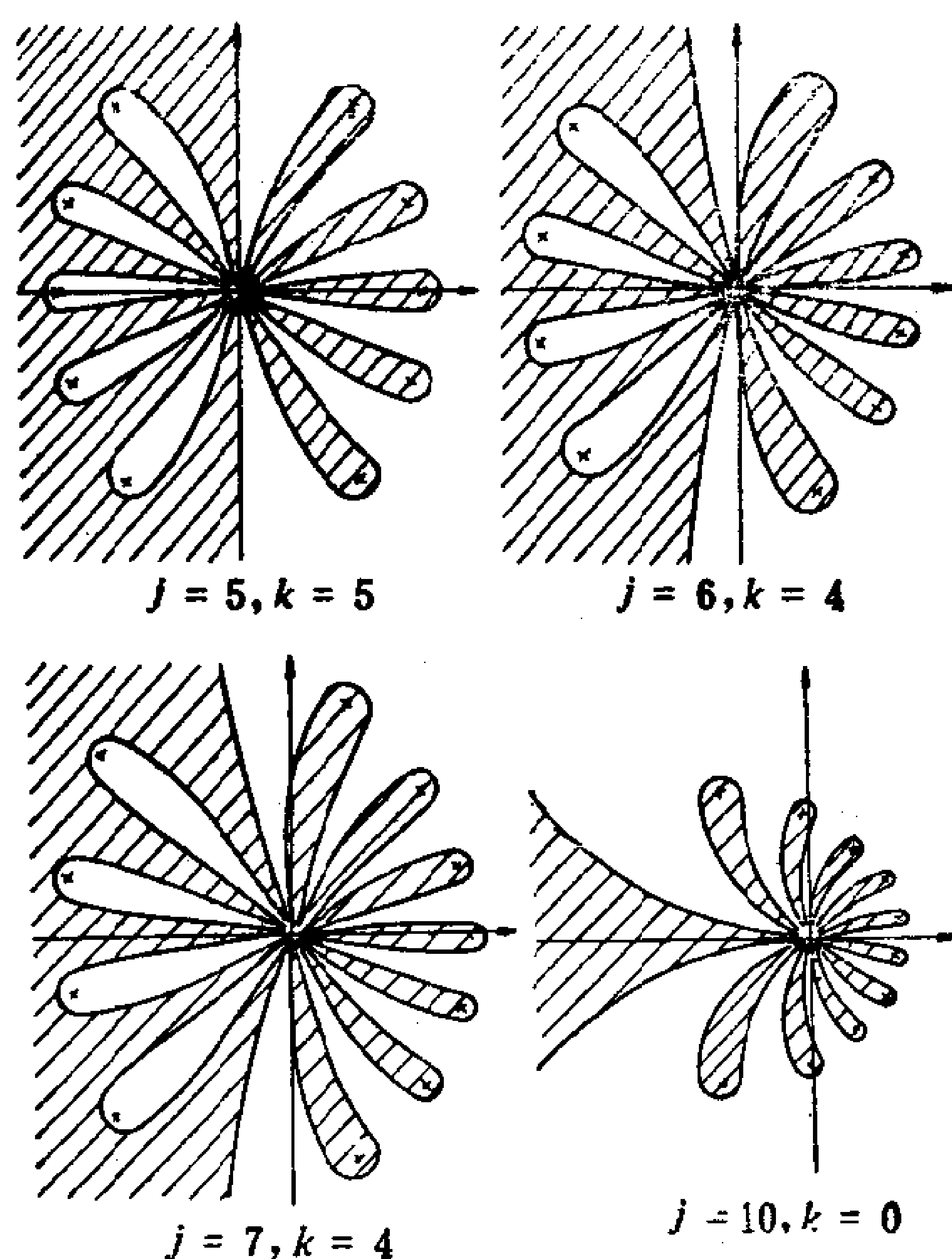


图 4.1 e^z 的 Padé 近似的 A 集合（阴影部分）

下面的几个定理是非常初等的，但对于讨论集合 A 起着基本的作用.

定理 4.3 设 $B_r = \{\theta \mid \theta \in [0, 2\pi],\ re^{i\theta} \in A\}$，于是存在数 r_0，使对 $r \geqslant r_0$，B_r 恰好是 $[0, 2\pi]$ 中的一个区间. 当 $r \to \infty$ 时，这个区间收敛到 $\left[\frac{\pi}{2}, \frac{3\pi}{2}\right]$. 所以集合 A 的边界 ∂A 只有二支趋向于无穷.

证明 由于 e^x 增长得比任何有理函数快，而 e^{-x} 减小得比任何有理函数快，对于固定的 t，有

$$\lim_{r\to\infty}\left|\frac{R(re^{it})}{e^{re^{it}}}\right| = \begin{cases} 0, & \text{如果} -\frac{\pi}{2} < t < \frac{\pi}{2}, \\ \infty, & \text{如果} \ \frac{\pi}{2} < t < \frac{3\pi}{2}, \end{cases}$$

因此，边界 ∂A 与圆 $z=re^{it}$ $(r>r_0)$ 至少有二个交点．为了证明最多有二个交点，我们在满足 $|R(re^{it})|=e^{r\cos t}$ 的 t 处，计算导数

$$\frac{d}{dt}\left((e^{r\cos t})^2-|R(re^{it})|^2\right)$$

$$=2re^{2r\cos t}\left(-\sin t-\mathrm{Re}\left(ie^{it}\frac{R'(re^{it})}{R(re^{it})}\right)\right).$$

由于当 $r\to\infty$ 时，$|R'/R|\to 0$，对于 $0<t<\pi$，当 r 充大时，这个导数将 <0；而对于 $\pi<t<2\pi$，这个导数将 >0．因此对于充分大的 r 只可能有二个交点．

下面的定理将集合 A 的形状与近似阶建立联系．由定义 4.1，所谓 $R(z)$ 为 e^z 的 p 阶近似是指存在常数 $c\neq 0$，有

$$e^z-R(z)=cz^{p+1}+O(z^{p+2})\quad(z\to 0). \tag{4.17}$$

定理 4.4 $R(z)$ 为 e^z 的 p 阶近似的充分必要条件是集合 A 由 $p+1$ 个宽度为 $\pi/(p+1)$ 的齿弧所组成，并且它们是由 $p+1$ 个具有同样宽度的 A 的补 $cA=\{z\in C\,|\,z\notin A\}$ 中的齿弧分开．

证明 $z=re^{it}$ 在 A 中的充要条件是 $|R(re^{it})e^{-r\cos t}|>1$．对于 $r\to 0$．由(4.17)得到条件

$$|1-ce^{-r\cos t}r^{p+1}e^{i(p+1)t}|>1.$$

由此导出

$$c\mathrm{Re}(e^{i(p+1)t})=c\cos(p+1)t<0,$$

这在逐次的 $(p+1)$ 个长度为 $\pi/(p+1)$ 的区间上是满足的．反之也然．定理证毕．

由于这个定理，称集合 A 为阶星形．另外称这些齿弧的每一个连通分量为指．如果 m 个齿弧联合成一个指，称它为重数是 m 的指．对于补 CA 中的类似的集合，我们分别称其为对偶指和重数是 m 的对偶指．

定理 4.5 每个重数是 m 的有界的指至少包含 $R(z)$ 的 m 个极点(并且计及它们的重数)．每个重数为 m 的有界的对偶指至少包含 $R(z)$ 的 m 个零点．

证明　令 $c(t), t_0 \leqslant t \leqslant t_1$，是指 F（见图 4.2）的正定向边界的参数化表示，$a = (c_1'(t), c_2'(t))$ 是切线向量，$n = (c_2'(t), -c_1'(t))$ 是外法线向量．记 $S(z) = r(x,y)e^{i\varphi(x,y)}$，$z = x + iy$，由于 $s(z)$ 的模往 F 内是增的，我们有 $\partial(\log r)/\partial n < 0$．现在极坐标形式的 Cauchy-Riemann 微分方程

$$\frac{\partial(\log r)}{\partial x} = \frac{\partial\varphi}{\partial y};\quad \frac{\partial(\log r)}{\partial y} = -\frac{\partial\varphi}{\partial x}$$

推出 $\partial\varphi/\partial a < 0$．因此 $s(z)$ 的幅角沿着 $c(t)$ 是减小的．在 $c(t)$ 的内部零点的数目 Z 和极点的数目 P 之间的差是

$$Z - P = \frac{1}{2\pi i}\int_c \frac{s'(z)}{s(z)}dz = s(z) \text{ 的幅角沿 } c(t) \text{ 旋转的次数}.$$

如果 F 是 m 重指，边界将 m 次回到原点，因此幅角 $\arg(s(z))$ 至少 m 次有同样的方向，所以旋转数至少为 $-m$ 和 $P \geqslant m$．（见图 4.2，其中 $m = 3$）

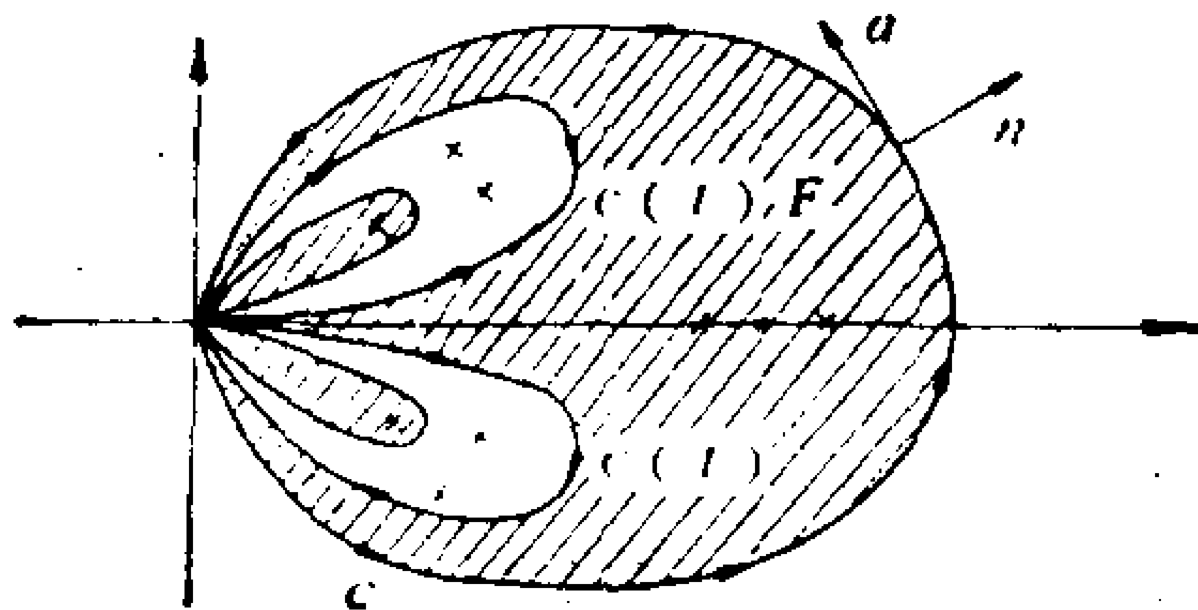


图 4.2

对于对偶指的讨论是同样的，可以从 $\dfrac{\partial(\log r)}{\partial n} > 0$ 开始讨论．

§2　e^z 的 Padé 近似的零点和极点

为验证定理 4.1 的条件 (iii)，需要确定有理函数 $R(z)$ 的极点在复平面上的分布．也就是要研究分母多项式的零点的分布问题．

我们先介绍一个较一般的关于多项式序列无零点区域的定理．

定理 4.6　令 $\{p_k(z)\}_{k=0}^n$ 是次数分别为 k 的多项式序列，满

足三项递推关系式

$$p_k(z) = \left(\frac{z}{b_k} + 1\right) p_{k-1}(z) - \frac{z}{c_k} p_{k-2}(z),$$

$$k = 1, 2, \cdots, n, \tag{4.18}$$

其中对于所有 $1 \leqslant k \leqslant n$, b_k 和 c_k 均是正实数，并且假设 $p_{-1}(z) \equiv 0$, $p_0(z) = p_0 \neq 0$. 令

$$\alpha = \min\{b_k(1 - b_{k-1}c_k^{-1}) \mid k = 1, 2, \cdots, n\}, \quad b_0 = 0. \tag{4.19}$$

如果 $\alpha > 0$, 则在抛物线区域

$$\mathscr{P}_\alpha = \{z = x + iy \in C \mid y^2 \leqslant 4\alpha(x + \alpha), x > -\alpha\} \tag{4.20}$$

中将不含 $p_1(z), p_2(z), \cdots, p_n(z)$ 的零点.

证明　令 $z \in \mathscr{P}_\alpha$ 为任意固定的点，并且不是任何 $p_k(z)$, $1 \leqslant k \leqslant n$ 的零点. 定义

$$\mu_k = \mu_k(z) = \frac{z p_{k-1}(z)}{b_k p_k(z)}, \quad k = 1,2,\cdots,n. \tag{4.21}$$

由归纳法,可以证明不等式

$$\operatorname{Re} \mu_k \leqslant 1, \quad k = 1, 2, \cdots, n. \tag{4.22}$$

这对于 $k = 1$ 是成立的. 事实上，由 (4.21), (4.18) 和假定 $p_0(z) = p_0 \neq 0$,有

$$\mu_1 = \frac{z p_0(z)}{b_1 p_1(z)} = \frac{z p_0(z)}{b_1(z/b_1 + 1) p_0(z)} = \frac{z}{z + b_1}.$$

由此推出,当且仅当 $\operatorname{Re} z \geqslant -b_1$ 时, $\operatorname{Re} \mu_1 \leqslant 1$. 因为 $z \in \mathscr{P}_\alpha$ 且由 (4.19)知 $b_1 \geqslant \alpha$,所以 $\operatorname{Re} z > -\alpha \geqslant -b_1$.

现在归纳地假定对某个满足 $2 \leqslant k \leqslant n$ 的 k,有 $\operatorname{Re} \mu_{k-1} \leqslant 1$. 由 (4.21) 和 (4.18), 可以将 μ_k 表成

$$\mu_k = \frac{z p_{k-1}(z)}{b_k p_k(z)} = \frac{z p_{k-1}(z)}{(z + b_k) p_{k-1}(z) - b_k c_k^{-1} z p_{k-2}(z)}$$

$$= \frac{z}{z + b_k - b_k c_k^{-1} b_{k-1} \mu_{k-1}}$$

或者写成 $\mu_k = T_k(\mu_{k-1})$,其中 $T_k(\omega)$ 是由

$$\xi = T_k(\omega) = \frac{z}{z + b_k - b_k c_k^{-1} b_{k-1} \omega}$$

所定义的变换．由归纳假设，μ_k 属于半平面 $\operatorname{Re}\omega\leqslant 1$ 在变换 $T_k(\omega)$ 下的象中．$T_k(\omega)$ 的唯一的极点是

$$\omega_k=\frac{z+b_k}{b_kc_k^{-1}b_{k-1}}.$$

再由(4.19)，$\operatorname{Re}z>-\alpha\geqslant-(b_k-b_kc_k^{-1}b_{k-1})$，可推得 $\operatorname{Re}\omega_k>1$．因此，$T_k(\omega)$ 将 $\operatorname{Re}\omega\leqslant 1$ 映到 ξ 平面中的闭圆盘 D_k 上．这个圆盘的中心 ξ_k 是在 $T_k(\omega)$ 下 ω 平面中相对于直线 $\operatorname{Re}\omega=1$ 与极点 ω_k 对称的点的象，即

$$\xi_k=T_k(2-\bar{\omega}_k)=T_k\left(2-\frac{\bar{z}+b_k}{b_kc_k^{-1}b_{k-1}}\right)$$

$$=\frac{z}{2\operatorname{Re}z+2b_k(1-b_{k-1}c_k^{-1})}.$$

另外，由于 $T_k(\infty)=0$ 在 D_k 的边界上，这个圆盘的半径 r_k 为

$$r_k=|\xi_k|=\frac{|z|}{2\operatorname{Re}z+2b_k(1-b_{k-1}c_k^{-1})},$$

因此，D_k 中的任意点的实部不超过和

$$\operatorname{Re}\xi_k+r_k=\frac{\operatorname{Re}z+|z|}{2\operatorname{Re}z+2b_k(1-b_{k-1}c_k^{-1})}.$$

再由(4.19)，这个量的上确界为

$$\frac{\operatorname{Re}z+|z|}{2\operatorname{Re}z+2\alpha},$$

可以直接验证，由于 $z\in\mathscr{P}\alpha$，它不会超过 1．于是由于 $\mu_k\in D_k$，则有 $\operatorname{Re}\mu_k\leqslant 1$．这就完成了建立(4.22)的归纳法．

容易看出，对于所有 $k=0,1,\cdots,n$，$p_k(0)\neq 0$．另外，如果对某个 $k\geqslant 1$，$p_k(z_0)=p_{k-1}(z_0)=0$，则显然有 $z_0\neq 0$，由(4.18) 导出对所有 $0\leqslant j\leqslant k$ 有 $p_{k-j}(z_0)=0$，特别这将推出 $p_0(z_0)=0$．得到矛盾．因此，对于每个 k，$1\leqslant k\leqslant n$，$p_k(z)$ 和 $p_{k-1}(z)$ 无公共的零点．

最后，假定定理不成立．存在某个 $z_0\in\mathscr{P}\alpha$ 和某个 k，$1\leqslant k\leqslant n$ 使有 $p_k(z_0)=0$．显然，由 (4.18)，$p_1(z)=(p_0(z)/b_1)(z+b_1)$．于是 p_1 的唯一零点是 $-b_1$．但由(4.19)，$-b_1\leqslant-\alpha$，这个

零点不在 $\mathscr{P}\alpha$ 中．因此，$2\leqslant k\leqslant n$．其次，由(4.18)，$p_k(z_0)=0$ 推出

$$\left(\frac{z_0}{b_k}+1\right)p_{k-1}(z_0)=\frac{z_0}{c_k}p_{k-2}(z_0),$$

并且因为 $p_k(z)$ 和 $p_{k-1}(z)$ 无公共的零点，通过除法得

$$\frac{c_k}{b_{k-1}b_k}(z_0+b_k)=\frac{z_0p_{k-2}(z_0)}{b_{k-1}p_{k-1}(z_0)}=\mu_{k-1}(z_0).$$

由(4.22)推出 $\mathrm{Re}\,\mu_{k-1}(z_0)\leqslant 1$．对上式取实部，得到

$$\mathrm{Re}\,z_0\leqslant -b_k(1-b_{k-1}c_k^{-1})\leqslant -\alpha,$$

另一方面，由(4.20)，$z_0\in\mathscr{P}\alpha$ 推出 $\mathrm{Re}\,z_0>-\alpha$，它与上面的不等式矛盾．因此对于每个 k，$1\leqslant k\leqslant n$，$p_k(z)$ 在 $\mathscr{P}\alpha$ 中无零点．证完．

定理 4.6 可以直接推广到满足 (4.18) 的无限多项式序列 $\{p_k(z)\}_{k=0}^{\infty}$．在这种情形，我们定义

$$\alpha=\inf\{b_k(1-b_{k-1}c_k^{-1})\mid k=1,\ 2,\ \cdots\},$$

于是只要 $\alpha>0$，则在由(4.20)定义的区域 $\mathscr{P}\alpha$ 中，每个 $p_k(z)$，$k=1,\ 2,\cdots$ 均无零点．

由这定理可得下面的一些推论．

推论 4.1 令 $s_k(z)=\sum_{j=0}^{k}a_jz^j$，$k=0,\ 1,\cdots,n$，假定对所有 $j=0,\ 1,\ \cdots,\ n$，$a_j>0$ 和

$$\alpha=\min\left\{\left(\frac{a_{k-1}}{a_k}-\frac{a_{k-2}}{a_{k-1}}\right)\middle|\,k=1,\ 2,\ \cdots,n\right\}>0,$$

其中约定 $a_{-1}/a_0=0$，于是多项式 $s_k(z)$，$k=1,\ 2,\ \cdots,\ n$ 在由(4.20)定义的抛物线区域 $\mathscr{P}\alpha$ 中无零点．

推论 4.2 令 $f(z)=\sum_{j=0}^{\infty}a_jz^j$ 是形式幂级数，假定对固定的 $\nu\geqslant 0$，对应的由(4.8)定义的 Hankel 行列式满足

$$A_k^{(\nu)}>0,\ A_k^{(\nu+1)}>0,\ \text{对}\ k=0,\ 1,\ \cdots,\ n,$$
$$A_k^{(\nu+2)}>0\ \text{对}\ k=0,\ 1,\ \cdots,\ n-1.$$

定义正常数 α 为

$$\alpha = \min\left\{\frac{A_k^{(\nu)}A_{k-1}^{(\nu+2)}}{A_{k-1}^{(\nu+1)}A_k^{(\nu+1)}}\,\middle|\, k = 1, 2, \cdots, n\right\}, \qquad (4.23)$$

于是 $f(z)$ 的 Padé 近似的分子 $N_{\nu,1}(z)$, $N_{\nu,2}(z)$, $\cdots$, $N_{\nu,n}(z)$ 在由(4.20)定义的抛物线区域 $\mathscr{P}\alpha$ 中无零点.

对于 Padé 近似的分母,可导出下面的类似结果.

推论 4.3 对于固定的 $n \geqslant 0$, 假定对应于形式幂级数 $f(z) = \sum_{j=0}^{\infty} a_j z^j$ 的 Hankel 行列式满足

$$A_n^{(k)} > 0,\ A_{n+1}^{(k)} > 0,\ 对于\ k = 1, 2, \cdots, \nu,$$
$$A_{n+2}^{(k)} > 0,\ 对于\ k = 1, 2, \cdots, \nu - 1.$$

由

$$\alpha = \min\left\{\frac{A_n^{(k)}A_{n+2}^{(k-1)}}{A_{n+1}^{(k-1)}A_{n+1}^{(k)}}\,\middle|\, k = 1, 2, \cdots, \nu\right\}$$

定义常数 α, 则对 $f(z)$ 的 Padé 近似的分母 $D_{1,n}(z)$ $D_{2,n}(z)$, $\cdots$, $D_{\nu,n}(z)$ 在抛物区域

$$\hat{\mathscr{P}}\alpha = \{z = x + iy \in C \mid y^2 \leqslant 4\alpha(\alpha - x),\ \alpha > x\}$$

中无零点.

这一节的其余部分考虑 e^z 的 Padé 近似. 由推论 4.2 和推论 4.3 可得到下面的推论.

推论 4.4 对于每个固定的 $\nu \geqslant 0$ 和每个 $n \geqslant 0$, e^z 的 Padé 近似的分子 $N_{\nu,n}(z)$ 在区域

$$\mathscr{P}_{\nu+1} = \{z = x + iy \in C \mid y^2 \leqslant 4(\nu + 1)(x + \nu + 1),$$
$$x > -(\nu + 1)\}$$

中无零点,而分母 $D_{\nu,n}(z)$ 在区域

$$\hat{\mathscr{P}}_{n+1} = \{z = x + iy \in C \mid y^2 \leqslant 4(n + 1)(n + 1 - x),$$
$$x < (n + 1)\}$$

中无零点.

特别, e^z 的 Padé 近似在区域

$$\mathscr{P}_1 = \{z = x + iy \in C \mid y^2 \leqslant 4(x + 1),\ x > -1\}$$

中无零点．

证明　对于 e^z．当 $S\geqslant 1$ 时，Hankel 行列式 $A_m^{(s)}$ 由

$$A_m^{(s)}=\prod_{j=1}^{s}\frac{1}{j(j+1)\cdots(j+m-1)}$$

给出．因此，对任何 $n\geqslant 0$，由(4.23)确定的常数为

$$\alpha=\min\{(\nu+1)\,|\,k=1,2,\cdots,n\}=\nu+1,$$

所以由推论 4.2，在区域 $\mathscr{P}_{\nu+1}$ 中 $N_{\nu,n}(z)$ 无零点．

对于 $D_{\nu,n}(z)$ 可进行类似的证明．

利用推论 4.4，通过繁琐但初等的证明可以得到下面一系列定理和推论，它们的证明见 Saff 和 Varga [100]，[101]．

定理 4.7　对于每个 $n\geqslant 2,\nu\geqslant 0$，$e^z$ 的 Padé 近似 $P_{\nu,n}(z)$ 在无限扇形

$$\varphi_{\nu,n}=\left\{z\in C\,\middle|\,|\arg z|\leqslant\cos^{-1}\left(\frac{n-\nu-2}{n+\nu}\right)\right\}\tag{4.24}$$

中无零点．

将这个定理应用到 $P_{[\sigma n],n}$，其中 $\sigma>0$，$[\sigma n]$ 表示不超过 σn 的最大的整数，得到下面的推论．

推论 4.5　对于任何固定的 $\sigma>0$，e^z 的 Padé 近似的序列 $\{P_{[\sigma n],n}(z)\}_{n=0}^{\infty}$ 在无限扇形

$$\varphi_\sigma=\left\{z\in C\,\middle|\,|\arg z|\leqslant\cos^{-1}\left(\frac{1-\sigma}{1+\sigma}\right)\right\}$$

中无零点．

由定理 4.7，如果 $n\leqslant\nu+2$，由(4.24)定义的无限扇形将含有整个闭右半平面．因此由定理 4.7，立即得到 Fhle [52] 的结果．

推论 4.6　如果 $n\leqslant\nu+2$，则 e^z 的 Padé 近似 $P_{\nu,n}(z)$ 的所有零点均在开左半平面．

这个推论的结果可以推广到 $n\leqslant\nu+4$ 的情形．

定理 4.8　如果 $n\leqslant\nu+4$，则 e^z 的 Padé 近似 $P_{\nu,n}(z)$ 的所有零点均在开左半平面中．

由于 $P_{0,5}(z)$ 在右半平面中具有零点，定理 4.8 中的结果是不能改进的．但是可以建立下面的结果．

定理 4.9 给定任何整数 τ，存在整数 $m=m(\tau)$，使 e^z 的 Padé 近似的序列 $\{P_{n-\tau,n}(z)\}_{n=m}^{\infty}$ 的所有零点均在开左半平面．

这个定理说明，沿着 Padé 表的任意 τ 次对角线充分远的 Padé 近似的所有零点均在左半平面中．

如果我们称使序列 $\{N_{n-\tau,n}(z)\}_{n=m}^{\infty}$ 在闭右半平面中无零点的最小的非负整数 m 为 $\tilde{m}(\tau)$，则数值计算表明有 $\tilde{m}(5)=6$，$\tilde{m}(6)=9$，$\tilde{m}(7)=14$，$\tilde{m}(8)=19$，$\tilde{m}(9)=26$．所以 $\tilde{m}(\tau)$ 似乎是 τ 的单调递增函数．

下面是两个关于包含 Padé 近似的所有零点的半平面和圆盘的结果．

定理 4.10 对于任何 $n\geqslant 1$ 和任何 $\nu\geqslant 0$，e^z 的 Padé 近似 $P_{\nu,n}(z)$ 的所有零点满足

$$\mathrm{Re}(z)<n-\nu.$$

定理 4.11 对于任何 $n\geqslant 3,\nu\geqslant 0$，$e^z$ 的 Padé 近似 $P_{\nu,n}(z)$ 的所有零点在圆盘

$$|z|=\frac{2(n+\nu)(n+\nu-1)}{(n+2\nu+1)}$$

中，另外，在 $\mathrm{Re}(z)\geqslant 0$ 中的 $P_{\nu,n}(z)$ 的零点均满足不等式

$$|z|\leqslant 2(n-3).$$

由(4.11)和(4.12)，成立等式

$$D_{i,k}(z)=N_{k,i}(-z), \tag{4.25}$$

因此，由前面 e^z 的 Padé 近似的零点的分布区域的结果，立即可以得到 Padé 近似的极点的分布区域的定理．例如，由定理 4.7，可以得到

定理 4.12 对于每个 $\nu\geqslant 2$，$n\geqslant 0$，e^z 的 Padé 近似 $P_{\nu,n}(z)$ 在无限扇形

$$\varphi_{\nu,n}=\left\{z\varepsilon C\,\middle|\,|\arg(-z)|\leqslant\cos^{-1}\left(\frac{\nu-n-2}{n+\nu}\right)\right\}$$

中无极点,因而 $P_{\nu,n}(z)$ 在 $\varphi_{\nu,n}$ 中是解析的.

由定理 4.8 可得到

定理 4.13 如果 $n\geqslant\nu-4$, 则 e^z 的 Padé 近似 $P_{\nu,n}(z)$ 的极点均在开右半平面中. 在开左半平面中 $P_{\nu,n}(z)$ 是解析的, 即 e^z 的 Padé 表中的第四次对角线及其上面的所有 Padé 近似在开左半平面中都是解析的.

§3 e^z 的有理近似在虚轴上的模

令

$$R_{j,k}(z)=\sum_{m=0}^{k}a_mz^m\Big/\left(\sum_{m=0}^{j}b_mz^m\right),a_0=b_0=1,b_j\neq 0 \quad (4.26)$$

是指数函数 e^z 的一个有理近似,有下面的定理.

定理 4.14 当 $k\leqslant j$ 时, (4.26)的有理函数 $R_{j,k}(z)$ 关于指数函数 e^z 的近似阶 $\geqslant k$ 的充分必要条件是 $R_{j,k}(z)$ 可唯一地表成

$$R_j^k(p;\ z)=\frac{\sum\limits_{m=0}^{k}(-1)^mp_j^{(j-m)}(0)z^m}{\left(\sum\limits_{m=0}^{j}(-1)^mp_j^{(j-m)}(1)z^m\right)}, \quad (4.27)$$

其中 $p_j(x)$ 是次数为 j 的多项式

$$p_j(x)=\sum_{m=0}^{j}r_m^jx^m,\quad p_j^{(r)}(a)=\frac{d^r}{dx^r}p_j(x)|_{x=a}, \quad (4.28)$$

并且有 $p_j^{(j)}(1)=1,\ p_j(1)\neq 0$.

证明 设 $R_{j,k}(z)$ 可唯一地表成(4.27). 由定义 4.1,我们必须证明

$$e^z\sum_{m=0}^{j}(-1)^mp_j^{(j-m)}(1)z^m$$

$$=\sum_{m=0}^{k}(-1)^mp_j^{(j-m)}(0)z^m+O(z^{k+1}), \quad (4.29)$$

应用两个级数的 Cauchy 乘积，得

$$e^z \sum_{m=0}^{j} (-1)^m p_i^{(j-m)}(1) z^m$$

$$= \sum_{m=0}^{\infty} \left\{ \sum_{l=0}^{\min(m,j)} \frac{(-1)^l}{(m-l)!} p_i^{(j-l)}(1) \right\} z^m,$$

但是，对于 $m \leqslant j$

$$\sum_{l=0}^{\min(m,j)} \frac{(-1)^l}{(m-l)!} p_i^{(j-l)}(1)$$

$$= (-1)^m \sum_{l=0}^{m} \frac{(-1)^l}{l!} p_i^{(j-m+l)}(1). \tag{4.30}$$

由 Taylor 展开，对于 $0 \leqslant i \leqslant j$，有

$$p_i^{(j-i)}(x) = \sum_{l=0}^{i} \frac{(x-1)^l}{l!} p_i^{(j-i+l)}(1). \tag{4.31}$$

因此(4.30)的右边等于 $(-1)^m p_i^{(j-m)}(0)$，(4.29) 成立.

假定 $R_{j,k}(z)$ 对 e^z 的近似阶 $\geqslant k$，由

$$p_i^{(j-m)}(1) = (-1)^m b_m,\ m = 0, 1, \cdots, j$$

确定唯一的 i 次多项式 $p_i(x)$，以及

$$e^z \sum_{m=0}^{j} (-1)^m p_i^{(j-m)}(1) z^m = \sum_{m=0}^{k} a_m z^m + O(z^{k+1}).$$

应用与上面充分性相类似的证明，可得到

$$a_m = (-1)^m p_i^{(j-m)}(0),\ m = 0, 1, \cdots, k,$$

这表示存在唯一的 i 次多项式 $p_i(x)$，使 $R_{j,k}(z)$ 可表成(4.27).
证完.

定义 4.4 令 $p_i(x)$ 是表示式(4.27)中的 x 的 i 次多项式，则称它为有理函数 $R_i^k(p,z)$ 的系数多项式(简称为 C 多项式).

定理 4.14 说明 $R_{j,k}(z)$ 对 e^z 的近似阶 $\geqslant k$ 的充分必要条件是存在唯一的 C 多项式 $p_i(x)$. 下面我们将下标省略掉，并且假定有 $k \leqslant j$.

如果我们知道近似的阶是大于 k 的，可以得到更加精细的结

果.

定理 4.15 令 $R_{j,k}(z)$ 是 e^z 的有理近似，则 $R_{j,k}(z)$ 的近似阶 $s \geqslant k$ 的充要条件是存在 $R_{j,k}(z)$ 的形式为

$$p(x) = \frac{d^{s-j}}{dx^{s-j}} [x^{s-k}(x-1)^{s-j} \hat{p}_{j,j+k-s}(x)], \text{ 如果 } s \geqslant j, \quad (4.32)$$

$$p^{(j-s)}(x) = x^{s-k} \hat{p}_{j,k}(x), \text{ 如果 } k \leqslant s \leqslant j \quad (4.33)$$

的唯一的 C 多项式，其中 $\hat{p}_{j,r}(x)$ 是 r 次多项式.

证明 首先假定 $R_{j,k}(z)$ 是 e^z 的阶 $s \geqslant k$ 的近似.由(4.29)有

$$\sum_{l=0}^{\min(m,j)} \frac{(-1)^l}{(m-l)!} p^{(j-l)}(1) = 0,$$

$$m = k+1, k+2, \cdots, s \quad (4.34)$$

(因为 $s \geqslant k$, C 多项式的存在性由定理 4.14 推出). 如果 $k < j$, 对于 $k+1 \leqslant m \leqslant j$, 由(4.30),(4.31)和(4.34)推出

$$p^{(j-m)}(0) = 0, \ m = k+1, k+2, \cdots, \min\{j, s\}, \quad (4.35)$$

于是对于 $s \leqslant j$ 得到(4.33). 现在假定 $s \geqslant j$, 定义函数 $I_r(x)$ $(r = 0, 1, 2, \cdots)$ 为

$$I_0(x) = p(x),$$
$$I_{r+1}(x) = \int_1^x I_r(t)dt, \ r \geqslant 0. \quad (4.36)$$

应用(4.31),我们得到显式

$$I_r(x) = \sum_{m=0}^{j} \frac{1}{(m+r)!} (x-1)^{m+r} p^{(m)}(1), \quad (4.37)$$

另外,由(4.34),对于 $m = j, j+1, \cdots, s$, 有

$$\sum_{l=0}^{j} \frac{(-1)^l}{(m-l)!} p^{(j-l)}(1)$$

$$= (-1)^j \sum_{l=0}^{j} \frac{(-1)^l}{(m-j+l)!} p^{(l)}(1) = 0. \quad (4.38)$$

由(4.35)(4.37)和(4.38), 推出

$$p^{(j-m)}(0) = 0, \ m = k+1, k+2, \cdots, j, \quad (4.39)$$

$$I_r(0)=0,\ r=1,\ 2,\ \cdots,\ s-j. \tag{4.40}$$

另外显然有

$$I_r(1)=0,\ r=1,\ 2,\ \cdots,\ s-j. \tag{4.41}$$

但是由 $I_r(x)$ 的定义(4.36),有

$$I'_{r+1}(x)=I_r(x),\ r\geqslant 0.$$

这与(4.39)—(4.41)一起给出

$$I_{s-j}^{(m)}(0)=0,\ m=0,\ 1,\cdots,\ s-k-1,$$
$$I_{s-j}^{(m)}(1)=0,\ m=0,\ 1,\ \cdots,\ s-j-1.$$

于是由于 $I_{s-j}(x)$ 是 s 次多项式,得

$$I_{s-j}(x)=x^{s-k}(x-1)^{s-j}\hat{p}_{j,j+k-s}(x), \tag{4.42}$$

其中 $\hat{p}_{j,j+k-s}(x)$ 是次数 $j+k-s$ 的唯一的多项式,但是

$$p(x)=I_{s-j}^{(s-j)}(x),$$

当 $s\geqslant j$ 时,这与(4.42)一起给出(4.32).

如果 C 多项式有形式(4.32)(4.33),则显然 $R_{t,k}(z)$ 的近似阶为 s. 证完.

令 $p(x)$ 是有理函数的 C 多项式,由

$$\tilde{p}(y)=2^n p\left(\frac{y+1}{2}\right) \tag{4.43}$$

定义另外一个多项式，我们称其为有理函数的 $\tilde{C}$ 多项式. 于是(4.27) 中的函数可以写成

$$\tilde{R}_j^k(\tilde{p};z)=\frac{\sum_{m=0}^{k}(-1)^m 2^{-m}\tilde{p}^{(j-m)}(-1)z^m}{\sum_{m=0}^{j}(-1)^m 2^{-m}\tilde{p}^{(j-m)}(1)z^m}, \tag{4.44}$$

并且 $\tilde{p}^{(j)}(1)=1$. 如果将(4.32)和(4.33)换成

$$\tilde{p}(y)=\frac{d^{s-j}}{dy^{s-j}}[(y+1)^{s-k}(y-1)^{s-j}\tilde{\tilde{p}}_{j,j+k-s}(y)],\ s\geqslant j, \tag{4.45}$$

$$\tilde{p}^{(j-s)}(y)=(y+1)^{s-k}\tilde{\tilde{p}}_{j,k}(y),\ k\leqslant s\leqslant j, \tag{4.46}$$

其中 $\tilde{\tilde{p}}_{j,r}(y)$ 是次数为 r 的多项式，则定理 4.15 的结果可以用 $\tilde{p}(y)$ 来叙述. 作这种变换的一个理由是为了对称性. 特别是对于 Padé 近似，Legendre 多项式将起重要的作用，通常这些多项

式都是定义在[-1, 1]上的。

例 4.15 e^x 的对角线和次对角线 Padé 近似的 $\tilde{C}$ 多项式由

$$\tilde{p}(y) = \frac{1}{(j+k)!} \frac{d^k}{dy^k} [(y+1)^j (y-1)^k] \qquad (4.47)$$

给出。

证明 由 Padé 近似的定义，第 (j, k)-Padé 近似的近似阶为 $s = j + k$. 将 s 的这个值代入(4.32)，再由 $\tilde{p}^{(j)}(x) = 1$，给出(4.47)。

利用 Legendre 多项式的正交性，可以证明下面的结论。

引理 4.1 令 $\tilde{p}(y)$ 由 (4.47) 给出，则存在唯一的常数组 $\tilde{a}_i$，$i = k, k+1, \cdots, j$，使有

$$\tilde{p}(y) = \sum_{i=k}^{j} \tilde{a}_i p_i(y), \qquad (4.48)$$

其中 $p_i(y)$ $(i = 0, 1, 2, \cdots)$ 是 Legendre 多项式。

利用表示式 (4.48)，得到下面的表示。

例 4.2 对于对角线，第一和第二次对角线的 Padé 近似的 $\tilde{C}$ 多项式分别为

$$\tilde{p}_k^0(y) = \frac{2^k k!}{(2k)!} p_k(y),$$

$$\tilde{p}_k^1(y) = \frac{2^{k-1}(k-1)!}{(2k-1)!} [p_{k-1}(y) + p_k(y)],$$

$$\begin{aligned} \tilde{p}_k^2(y) &= \frac{2^{k-1}(k-2)!}{(2k-1)!} [k p_{k-2}(y) \\ &\quad + (2k-1)p_{k-1}(y) + (k-1)p_k(y)] \\ &= \frac{2^{k-1}(k-2)!}{(k-1)(2k-2)!} [(ky + k - 1)p_{k-1}(y) \\ &\quad - p_k(y)]. \end{aligned}$$

现在我们考虑二个重要的有理分式近似类

例 4.3 $\tilde{C}$ 多项式 $\tilde{p}^{(0)}$ 给出阶 $2k$ 的对角线 Padé 近似，而 $\tilde{p}^{(1)}$ 给出阶为 $2k-1$ 的第一次对角线的 Padé 近似。我们可以研究包含这两个近似作为特例的所有阶至少为 $2k-1$ 的有理函数近似类。按定理 4.15 和方程 (4.44)，(4.45)，这类近似可以唯一地

写成

$$S_k(a;\ z)=\frac{\sum_{m=0}^{k}(-1)^m 2^{-m}\tilde{p}^{(k-m)}(-1)z^m}{\sum_{m=0}^{k}(-1)^m 2^{-m}\tilde{p}^{(k-m)}(1)z^m}, \tag{4.49}$$

其中 $\tilde{p}(y)$ 含有参数 a,

$$\tilde{p}(y)=PI_k(a;\ y)=\frac{1}{(2k-1)!}$$

$$\times\frac{d^{k-1}}{dy^{k-1}}[(y^2-1)^{k-1}(y+a)], \tag{4.50}$$

记这个多项式为 $PI_k(a;y)$. 由于

$$\frac{d^k}{dy^k}[(y^2-1)^k]=2k\frac{d^{k-1}}{dy^{k-1}}[(y^2-1)^{k-1}y]$$

$$PI_k(a;\ y)=\tilde{p}_k^0(y)+\frac{a}{2k-1}\tilde{p}_{k-1}^0(y)$$

$$=\frac{2^k k!}{(2k)!}[p_k(y)+ap_{k-1}(y)], \tag{4.51}$$

对于 $k=1,\ 2$, 有

$$PI_1(a;\ y)=y+a,$$

$$S_1(a;\ z)=\frac{1+\frac{1}{2}(1-a)z}{1-\frac{1}{2}(1+a)z},$$

$$PI_2(a;\ y)=\frac{1}{2}y^2+\frac{a}{3}y-\frac{1}{6},$$

$$S_2(a;z)=\frac{1+\frac{1}{2}\left(1-\frac{a}{3}\right)z+\frac{1}{12}(1-a)z^2}{1-\frac{1}{2}\left(1+\frac{a}{3}\right)z+\frac{1}{12}(1+a)z^2}.$$

由引理 4.1, 方程(4.51)和 $\tilde{C}$ 多项式的唯一性, 如果 $a=0$, 则近似(4.49)的近似阶为 $2k$. 在这种情形我们得到 Padé 近似. 对于 $a\neq 0$, 近似阶为 $2k-1$, 当 $a=1$ 时产生第一次对角线 Padé

近似。

例 4.4　阶大于或等于 $2k-2$ 的有理分式近似族。按 (4.45)，$\tilde{C}$ 多项式 $\tilde{p}(y)$ 有形式

$$\tilde{p}(y)=\frac{1}{(2k-2)!}\times\frac{d^{k-2}}{dy^{k-2}}[(y^2-1)^{k-2}(y^2+ay+b)]. \quad (4.52)$$

与例 4.3 一样

$$\frac{d^k}{dy^k}[(y^2-1)^k]=2k\frac{d^{k-1}}{dy^{k-1}}[(y^2-1)^{k-1}y]$$
$$=2k(2k-1)\frac{d^{k-2}}{dy^{k-2}}[(y^2-1)^{k-2}y^2]$$
$$-2k\frac{d^{k-2}}{dy^{k-2}}[(y^2-1)^{k-2}],$$

记这个 $\tilde{p}(y)$ 为 $PJ_k(a, b; y)$，于是 PJ_k 可写成为

$$PJ_k(a, b; y)=\tilde{p}_k^0(y)+\frac{a}{2(k-1)}\tilde{p}_{k-1}^0(y)+\frac{1}{2(k-1)(2k-3)}\left[b+\frac{1}{2k-1}\right]\tilde{p}_{k-2}^0(y)$$
$$=\frac{2^{k-2}(k-2)!}{(2k-2)!}\left[\frac{2(k-1)}{2k-1}p_k(y)+ap_{k-1}(y)+\left(b+\frac{1}{2k-1}\right)p_{k-2}(y)\right], \quad (4.53)$$

再由 $\tilde{C}$ 多项式的唯一性推出：如果 $b=-1/(2k-1)$，a 任意，则阶 $s\geqslant 2k-1$；如果 $b=-1/(2k-1)$，$a=0$，则阶 $s=2k$。将对应于 $PJ_k(a,b; y)$ 的有理分式记成 $T_k(a, b; z)$。前两个 $\tilde{C}$ 多项式和有理分式近似为

$$PJ_2(a, b; y)=\frac{1}{2}(y^2+ay+b),$$

$$T_2(a, b; z)=\frac{1+\frac{1}{2}\left(1-\frac{a}{2}\right)z+\frac{1}{4}\left(\frac{1+b}{2}-\frac{a}{2}\right)z^2}{1-\frac{1}{2}\left(1+\frac{a}{2}\right)z+\frac{1}{4}\left(\frac{1+b}{2}+\frac{a}{2}\right)z^2},$$

$$PJ_3(a, b; y) = \frac{1}{6}y^3 + \frac{1}{8}ay^2 + \frac{1}{12}(b-1)y - \frac{1}{24}a,$$

$$T_3(a, b; z) = \left[1 + \frac{1}{2}\left(1 - \frac{a}{4}\right)z + \frac{1}{16}\left(\frac{5+b}{3} - a\right)z^2 + \frac{1}{96}(1+b-a)z^3\right] \Big/ \left[1 - \frac{1}{2}\left(1 + \frac{a}{4}\right)z + \frac{1}{16}\left(\frac{5+b}{3} + a\right)z^2 - \frac{1}{96}(1+b+a)z^3\right].$$

利用 C 多项式 $p(x)$ 可以建立验证定理 4.1 的条件 (4.13) 的准则. 对于 $y \in R$, 定义函数

$$E_{j,k}(y) = |D_{j,k}(iy)|^2 - |N_{j,k}(iy)|^2. \tag{4.54}$$

条件(4.13)等价于

$$E_{j,k}(y) \geqslant 0, \quad y \in R.$$

对于 $k \leqslant j$, 定义函数 $H_j^k(\tau, v)$ 和 $M_j^k(\tau, y)$ 为

$$H_j^k(\tau, v) = \sum_{m=0}^{k} (-1)^m p^{(j-m)}(\tau) v^m, \quad v \in C, \quad \tau \in R, \tag{4.55}$$

$$M_j^k(\tau, y) = H_j^k(\tau, iy) \cdot H_j^k(\tau, -iy), \quad y \in R, \tag{4.56}$$

则有

$$|D_{j,k}(iy)|^2 = M_j^j(1, y),$$

$$|N_{j,k}(iy)|^2 = M_j^k(0, y).$$

由于 $M_j^k(\tau, y) = M_j^k(\tau, -y)$, 可以将其表成

$$M_j^k(\tau, y) = \sum_{m=0}^{k} \Gamma_{2m}^{j,k}(\tau) y^{2m}. \tag{4.57}$$

下面来确定 $\Gamma_{2m}^{j,k}(\tau)$ 的表达式. 直接对(4.55)微分,得到

$$\frac{\partial H_j^k(\tau, v)}{\partial \tau} = -vH_j^k(\tau, v) + (-1)^k p^{(j-k)}(\tau) v^{k+1}.$$

因此由(4.56)知

$$\frac{\partial M_j^k(\tau, y)}{\partial \tau} = 2(-1)^k p^{(j-k)}(\tau) \mathrm{Re}\{(-iy)^{k+1} H_j^k(\tau, iy)\}$$

$$= -2p^{(j-k)}(\tau) \mathrm{Re}\left\{\sum_{m=0}^{k} (-1)^m p^{(j-m)}(\tau)(iy)^{k+m-1}\right\},$$

但是由(4.57)

$$\frac{\partial M_j^k(\tau,y)}{\partial\tau}=\sum_{m=0}^{k}\frac{d}{d\tau}\Gamma_{2m}^{j,k}(\tau)y^{2m},$$

比较系数给出

$$\frac{d}{d\tau}\Gamma_{2m}^{j,k}(\tau)=\begin{cases}0, & \text{对于 } 2m\leqslant k,\\ 2(-1)^{k+m}p^{(j-k)}(\tau)\\ \quad\times p^{(j+k+1-2m)}(\tau), & \text{对于 } 2m>k.\end{cases}$$

通过积分得到

$$\Gamma_{2m}^{j,k}(\tau)=\begin{cases}\Gamma_{2m}^{j,k}(0), & \text{对于 } 2m\leqslant k,\\ \Gamma_{2m}^{j,k}(0)+2(-1)^{k+m}\int_0^\tau p^{(j-k)}(x)\\ \quad\times p^{(j+k+1-2m)}(x)dx, & \text{对于 } 2m>k.\end{cases}$$

假定有理函数 $R_{j,k}(z)$ 对 e^z 的近似阶 $s\geqslant j$，则对于 $m=k+1,k+2,\cdots,j$，$p^{(j-m)}(0)=0$. 这样，由(4.55)，(4.56)和(4.57)可以推得

$$\Gamma_{2m}^{j,k}(0)=\Gamma_{2m}^{j,j}(0),\quad \text{当 } m\leqslant k \text{ 时},$$
$$\Gamma_{2m}^{j,j}(0)=0,\quad \text{当 } m>k \text{ 时}.$$

将 $\Gamma_{2m}^{j,k}(\tau)$ 代入(4.57)，并记

$$\begin{aligned}E(p,j,k,y)&=M_j^j(1,y)-M_j^k(0,y)\\&=2(-1)^j\sum_{m=\left[\frac{(j+2)}{2}\right]}^{j-1}\left\{(-1)^m\int_0^1 p(t)p^{(2j-2m+1)}(t)\,dt\right\}y^{2m}\\&\quad+\{[p(1)]^2-[p(0)]^2\}y^{2j},\end{aligned}\tag{4.58}$$

则我们得到下面的定理

定理 4.16 令 $R_{j,k}(z)=R_j^k(p;z)$ 是近似阶 $s\geqslant j$ 的 e^z 的有理近似，则定理 4.1 中的条件 (i) 成立的充分必要条件为

$$E(p,j,k;y)\geqslant 0,\quad y\in R.\tag{4.59}$$

§4 A 可 接 受 性

由定理 4.1 和定理 4.16，直接得到定理

定理 4.17 令 $R_{j,k}(z)=R_j^k(p,z)$ 是 e^z 的近似阶 $s\geqslant j$ 的有理近似,则如果满足条件

(i) $E(p,j,k;y)\geqslant 0,\ y\in R,$ (4.60)

其中 E 定义为

$$E(p,j,k;y)=2(-1)^s\times\sum_{m=\left[\frac{(s+2)}{2}\right]}^{j-1}\left\{(-1)^m\int_0^1 I_{s-j}(x)p^{(j+s+1-2m)}(x)\,dx\right\}y^{2m}+\{[p(1)]^2-[p(0)]^2\}y^{2j},\tag{4.61}$$

I_r 由(4.36)给出.

(ii) $[p(1)]^2-[p(0)]^2\geqslant 0$.

(iii) 多项式

$$D(p,j;z)=\sum_{m=0}^{j}(-1)^m p^{(j-m)}(1)z^m$$

的所有零点均在右半复平面中.

则有理近似 $R_{j,k}(z)$ 是 A 可接受的.

证明 (ii) 和 (iii) 是定理 4.1 的条件 (ii) 和 (iii) 的直接应用. 只要证明(4.58) 可以写成 (4.61) 的形式, 则 (i) 是定理 4.16 的结果. 由定理的假定,阶 $s\geqslant j$, 这表示

$$I_r(0)=0,\ r=1,2,\cdots,s-j,$$

应用这个结果,并对(4.58) 进行若干次分部积分,得到 (4.61). 证完.

与 §3 中一样,也可以将定理中的区间[0,1]改变成[−1,1].下面只给出用 $\tilde{p}(y)$ 表示的函数 $\tilde{E}$. 应用(4.44),得到

$$\tilde{E}(\tilde{p},j,k;y)=2(-1)^s\times\sum_{m=\left[\frac{(s+2)}{2}\right]}^{j-1}\left\{(-1)^m 2^{-2m}\int_{-1}^1 \tilde{I}_{s-j}(x)\tilde{p}^{(j+s+1-2m)}(x)dx\right\}y^{2m}+2^{-2j}\{[\tilde{p}(1)]^2-[\tilde{p}(-1)]^2\}y^{2j}.\tag{4.62}$$

其中定义

$$\hat{I}_0(y) = \tilde{p}(y),$$

$$I_{r+1}(y) = \int_t^y \tilde{I}_r(t)dt,\ r \geqslant 0. \tag{4.63}$$

在例 4.5 中，将应用下面的一个定理

定理 4.18 如果多项式 $p(x)$（或 $\tilde{p}(y)$）满足

$$p(x) = (-1)^j p(1-x), \tag{4.64}$$

$$(\tilde{p}(y) = (-1)^j \tilde{p}(-y)).$$

则当 $s \geqslant j$ 时，相应的多项式 E（或 $\tilde{E}$）是零多项式.

证明 只考虑多项式 E 和 $j = k$ 的情形. 由(4.64)得

$$H_i^j(0, v) = H_i^j(1, -v).$$

因此有

$$M_i^j(0, y) = M_i^j(1, y),$$

这时多项式 E 可表成

$$E(p, j, j; y) = M_i^j(1, y) - M_i^j(0, y),$$

立即得到定理的结论.

例 4.5 对角线，第一下次对角线和第二下次对角线的 e^x 的 Padé 近似是 A 可接受的.

由定理 4.13 知这些 Padé 近似在开左半平面中是解析的. 定理 4.1 的条件 (ii),(iii) 显然满足.

对于对角线的 Padé 近似，$\tilde{p}_k^0(y)$ 满足(4.64)，由定理 4.18 知定理 4.17 的 (i) 成立. 对于第一下次对角线，由(4.62)，有

$$\tilde{E}(\tilde{p}^1, j, j-1; y) = \left[\frac{(j-1)!}{(2j-1)!}\right]^2 y^{2j},$$

因而(4.60)成立. 第二下次对角线的 Padé 近似的多项式为

$$\tilde{E}(\tilde{p}^2, j, j-2; y) = \left[\frac{(j-2)!}{(2j-2)!}\right]^2 y^{2j},$$

(4.60) 也成立. 因此上述的 A 可接受性的结论成立.

例 4.6 对于第三下次对角线 Padé 近似 $P_{j,k}(z)$，这时 $k = j-3$, $s = 2j-3$，多项式为

$$\tilde{E}(\tilde{p}^3, j, j-3; y) = \left[\frac{(j-3)!}{(2j-3)!}\right]^2 y^{(2j-2)}\{y^2 - j(j-2)\},$$

当 y 满足不等式

$$-\sqrt{j(j-2)} < y < \sqrt{j(j-2)}$$

时，(4.60)不成立，对于 $z = iy$，其对应的近似按模大于 1. 因此第三下次对角线 Padé 近似不是 A 可接受的.

正是由于例 4.5 和例 4.6 的结果，Ehle (1969) 提出下面的猜测

定理 4.19 (Ehle 猜测) 指数函数 e^z 的 Padé 近似 $P_{j,k}(z)$ 为 A 可接受的充分必要条件是

$$j-2 \leqslant k \leqslant j. \tag{4.65}$$

为了证明这个定理，需要以下二个定理. 它们的证明都是采用阶星形的技巧.

定理 4.20 如果 $R_{j,k}(z)$ 是 A 接受的，并且是 e^z 的阶为 s 的近似，则有

$$s \leqslant 2j \text{ 和 } s \leqslant 2k+2.$$

证明 由定理 4.4，阶星形 A 至少有 $[(s+1)/2]$ 个指出现在左半平面 C^- 中 (见图 4.3，其中 $s+1=11$.). 由定理 4.2 和定理 4.5，这些指不能与虚轴相交，也不能是有界的. 由定理 4.3，它们将合成一个指，并且至少包含 $[(s+1)/2]-1$ 个有界的对偶指. 而由定理 4.5，$R_{j,k}(z)$ 的零点的总数 k 满足不等式 $k \geqslant [(s+1)/2]-1$，即 $s \leqslant 2k+2$.

另外的不等式是显然的. 因为 $s \leqslant k+j$，和 $k \leqslant j$.

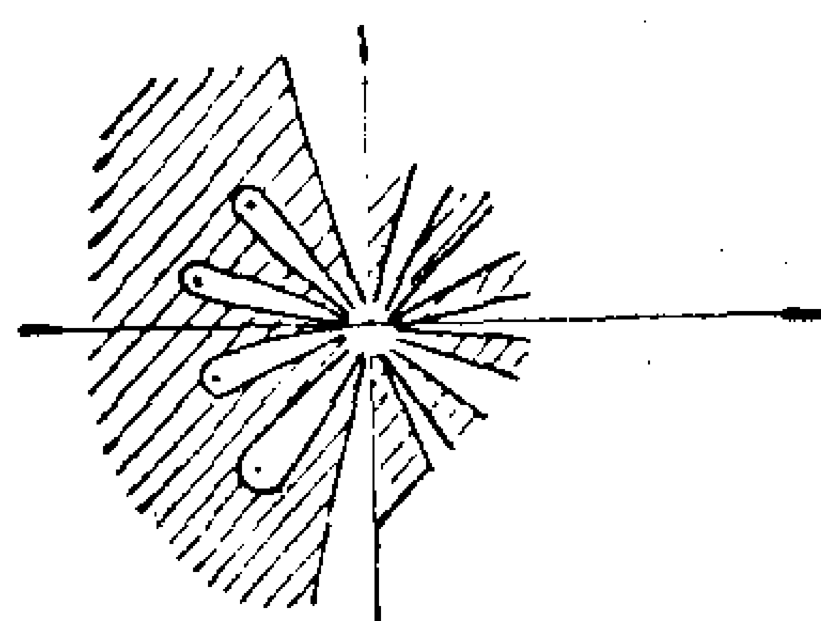


图 4.3

定理 4.21 假定 e^z 的有理近似 $R_{j,k}(z)$ 满足条件

(i) 近似阶 $s \geqslant 2j-2$,

(ii) $\lim\limits_{z\to\infty}|R_{j,k}(z)| \leqslant 1$,

(iii) $R_{j,k}(z)$ 的分母 $D_{j,k}(z)$ 的系数具有交替的符号。

则 $R_{j,k}(z)$ 是A可接受的.

证明 由 (ii) 推出 $k \leqslant j$. 由(4.54)定义的函数 $E_{j,k}(y)$ 是偶函数. 由阶的定义, $E_{j,k}(y) = O(y^{s+1})$, $y \to 0$. 因此, (i) 给出 $E_{j,k}(y) = Dy^{2j}$, 而由 (ii) 推出 $D \geqslant 0$. 阶星形不会与虚轴相交.

由定理 4.4, 至少有 $\left[\frac{(s+1)}{2}\right]$ 个A的指出现在右半平面中, 并且由定理 4.3, 它们一定是有界的. 因此 $R_{j,k}(z)$ 在右半平面中至少有 $\left[\frac{(s+1)}{2}\right]$ 个极点.

由 (i) $\left[\frac{s+1}{2}\right] \geqslant j-1$, $R_{j,k}(z)$ 在左半平面中最多有一个极点,它一定是实的. 但是由(iii), 这是不可能的. 这样定理 4.1 的三个条件都能满足, $R_{j,k}(z)$ 是A可接受的.

定理 4.19 的证明 必要性 由于 Padé 近似具有最优阶 $s = k+j$. 由定理 4.20, 推出

$$k+j \leqslant 2j, \quad k+j \leqslant 2k+2,$$

这就是 (4.65). 充分性由定理 4.21 或例 4.5 给出.

本 章 附 注

§1 的材料取自 Ralston [96], lambert[67] 和 Wanner, Hairer, Nørsett 的 [110].

§2 是根据 Saff, Varga 的 [100][101] 编写的.

§3 是根据 Nørsett [91] 编写的.

§4 的材料取自 Nørsett [91] 和 Wanner, Hairer, Nørsett [110].

第五章　指数拟合方法

考虑刚性常微分方程组

$$y' = f(t, y), \tag{5.1}$$

它由N个一阶方程所组成．它的时间常数是 Jacobi 矩阵 $J=J(t)=\frac{\partial f}{\partial y}(t, F(t))$ 的特征值的实部的负倒数，其中 $F(t)$ 是 (5.1) 的要求的特解．为了在 $y=F(t)$ 的邻域中考虑 (5.1) 的性态，可以考虑 (5.1) 的变分方程

$$y' = J(t)y \tag{5.2}$$

和它的特解． 假定 $J(t)$ 是慢变化的，(5.2) 的特征解近似为指数函数，即只要 λ 是 J 的单根，则

$$y(t) = y(0)e^{\lambda t} \tag{5.3}$$

将是 (5.2) 的一个近似的特征解．在格点 $\{t_n = nh\}$, $h>0$, $n=0,1,\cdots$上，解 (5.3) 满足递推式

$$y(t_{n+1}) = e^{q}y(t_n), \tag{5.4}$$

其中 $q=\lambda h$．因此当 $|q|\ll 1$ 时，由 $y(t_n)$ 到 $y(t_{n+1})$ 的变化是缓慢的． 大多数传统的积分公式只在这种情形是精确的． 若 $|q|\gg 1$，由 (5.4)，从 $y(t_n)$ 到 $y(t_{n+1})$ 的变化是迅速的．传统的数值积分公式无法反映这种变化．一个能有效地积分求解刚性方程组的数值积分公式对于 $|q|\ll 1$ 和对于所有或至少一些指定的任意固定的大的 $-\mathrm{Re}q$ 值应该是精确的和稳定的．为了对于大的 q 达到精确的要求，一些作者提出直接考虑应用迅速衰减的指数函数来构造解，而不去应用解的 Taylor 展开． 这一章采用这个观点． 精确地说，我们将考虑一些含自由参数(除 h 外)的积分公式族，选取这些公式中的参数，使得这些公式对于刚性方程组的某些迅速变化的解是精确的．称这样构造的方法为指数拟合方法．

指数拟合方法的稳定区域是在 $h\lambda$ 平面上进行指数拟合的点的邻域的和. 在一些情形中,稳定区域可以是无限的,例如整个左半平面. 指数拟合方法比起其它许多适合刚性问题的方法(例如许多 A 稳定方法)有一个优点,即如果拟合得适当和暂态问题可以近似线性化时,可以用大步长积分,当暂态是高频振荡,而且衰减得非常慢时,这个性质是有价值的.

§1 指数拟合方法

对于指数拟合的概念,我们引进下面的定义

定义 5.1 一个数值方法称作在(复)值 λ_0 处是指数拟合的,如果当方法以精确的初始条件应用到 $\lambda=\lambda_0$ 的数值试验问题

$$y'=\lambda y,\ y(0)=y_0 \tag{5.5}$$

时,将得到精确的理论解.

我们将讨论一些具体的算法来说明这个概念.

例 5.1 考虑恒等式

$$y(t+h)-y(t)-h[(1-\mu)y'(t+h)+\mu y'(t)]\equiv e_1(t), \tag{5.6}$$

其中 $y(t)$ 是二次连续可微的, μ 是实参数, h 为固定的正数,

$$e_1(t)=-h^2\int_0^1(\theta-\mu)y''(t+\theta h)d\theta, \tag{5.7}$$

由 (5.6) 构造积分公式

$$F^{(1)}:\ y_{n+1}-y_n-h[(1-\mu)y'_{n+1}+\mu y'_n]=0, \tag{5.8}$$

局部截断误差由 (5.7) 给出. 对 (5.7) 进行分部积分,可以看出,当 $\mu\neq\frac{1}{2}$ 时, $F^{(1)}$ 的精确阶为 $p=1$. 令 (5.8) 中的 $\mu=1$,得到向前的 Euler 公式. 而当 $\mu=0$ 时,得到向后的 Euler 公式. 对于 $\mu=\frac{1}{2}$ 和 $y(t)$ 是三次连续可微的,我们得到梯形法,这时 $p=2$,误差项为

$$\frac{h^3}{4}\int_0^1 2\theta(\theta-1)y^{(3)}(t+\theta h)d\theta. \tag{5.9}$$

将公式 (5.8) 应用到 (5.5)，得到的数值解满足递推式

$$y_{n+1} = r^{(1)}(q)y_n, \tag{5.10}$$

其中

$$r^{(1)}(q) = (1+\mu q)/(1-(1-\mu)q). \tag{5.11}$$

显然，公式 (5.8) 为 A 稳定的充分必要条件是对所有 q, $\mathrm{Re}q < 0$, 有

$$|r^{(1)}(q)| < 1.$$

由第二章，这个条件等价于

(i) 在 $\mathrm{Re}q = 0$ 上，$|r^{(1)}(q)| \leqslant 1$.

(ii) 在 $\mathrm{Re}q < 0$ 中，$r^{(1)}(q)$ 是解析的.

因而立即推出，当且仅当 $\mu \leqslant \frac{1}{2}$ 时，$F^{(1)}$ 是 A 稳定的.

(5.5) 的精确解具有递推式 (5.4)，为了求得指数拟合的参数，比较 (5.4) 和 (5.10)，应该考虑用 $r^{(1)}(q)$ 近似 e^q 的误差

$$\varepsilon^{(1)}(q) = r^{(1)}(q) - e^q. \tag{5.12}$$

如果对某些特殊的值 $q_0 = \lambda_0 h$，有 $\varepsilon^{(1)}(q_0) = 0$，则在离散的意义下，用 $F^{(1)}$ 计算得到的 $\lambda = \lambda_0$ 时问题 (5.5) 的数值解是精确的. 即对于所有 n 和任意固定的 $h > 0$，只要 $y_0 = y(0)$，就有 $y_n = y(t_n)$. 在定义 5.1 的意义下，公式 $F^{(1)}$ 在 $q = q_0$ 处是指数拟合的.

为了在拟合的量级上给出指数拟合的特征，需要拟合阶的概念. 设用某种数值方法求解 (5.5) 得到的解序列 $\{y_n\}_{n=0,1,\cdots}$ 满足递推式

$$y_{n+1} = r(q)y_n, \tag{5.13}$$

令

$$\varepsilon(q) = r(q) - e^q. \tag{5.14}$$

定义 5.2 一个方法称作在 $q = q_0 \neq 0$ 处为 ϕ 阶指数拟合的，如果 q_0 是 $\varepsilon(q)$ 的 $\phi+1$ 重零点.

这个定义表示，如果一个方法在 q_0 处是 ϕ 阶指数拟合的，则 q_0 是由 $r(q)$ 和 e^q 所表示的曲线之间的 ϕ 阶接触点．下面认为在 $q_0=0$ 处指数拟合到 p 阶是指在通常意义下 p 阶的多项式拟合．

对于 μ 的任意值，公式 $F^{(1)}$ 在 $q_0=0$ 处均拟合到 $p\geqslant 1$ 阶．另外，可以选取自由参数 μ，使得这个公式在别处也是拟合的．例如，令 $\mu=0$，我们得到向后 Euler 公式．它在 $q_0=0$ 处是 $p=1$ 阶拟合的，而在 $q_0=\infty$ 处是 $\phi=0$ 阶拟合的．如果我们不在 $q_0\neq 0$ 处进行指数拟合，而按传统的方式选取参数使在 $q_0=0$ 处的拟合阶 p 达到极大值，则对于 $|q|$ 大的值，用 $r^{(1)}(q)$ 来近似 e^q 可以是十分坏的．例如，在梯形法 $\left(\mu=\frac{1}{2}\right)$ 的情形，$F^{(1)}$ 在 $q_0=0$ 处拟合到最大阶 $p=2$，但是有

$$\lim_{q\to-\infty}\varepsilon^{(1)}(q)=\lim_{q\to-\infty}r^{(1)}(q)=-1,$$

即 $F^{(1)}$ 将任意地接近 A 稳定性的极限状态．

要求 $F^{(1)}$ 在给定的 $q_0(\phi=0$ 阶$)$拟合，将得到一个确定 μ 的方程，它的解为

$$\mu=\mu^{(1)}(q_0)=-q_0^{-1}-e^{q_0}(1-e^{q_0})^{-1}. \tag{5.15}$$

因此，对于任意给定的实数 λ_0，令 μ 取实值 $\mu^{(1)}(q_0)$，$q_0=\lambda_0 h$，$F^{(1)}$ 可以在 q_0 处指数拟合．记在 q_0 处指数拟合的公式 $F^{(1)}$ 为 $F^{(1)}[q_0]$．于是梯形法 $\left(\mu=\frac{1}{2}\right)$ 和向后 Euler 公式 $(\mu=0)$ 可以分别表成 $F^{(1)}[0]$ 和 $F^{(1)}[\infty]$．

例 5.2　考虑恒等式

$$y(t+h)-y(t)-\frac{h}{2}[(1+a)y'(t+h)+(1-a)y'(t)]$$
$$+\frac{h^2}{4}[(b+a)y''(t+h)-(b-a)y''(t)]\equiv e_2(t), \tag{5.16}$$

其中 $y(t)$ 三次连续可微，a、b 是实参数，

$$e_2(t)=\frac{h^3}{4}\int_0^1[2\theta^2-2(1-a)\theta$$

$$+ (b-a)]y^{(3)}(t+\theta h)d\theta. \tag{5.17}$$

由 (5.16)，构造积分公式

$$F^{(2)}: \; y_{n+1} - y_n - \frac{h}{2}[(1+a)y'_{n+1} + (1-a)y'_n]$$

$$+ \frac{h^2}{4}[(b+a)y''_{n+1} - (b-a)y''_n] = 0, \tag{5.18}$$

其局部截断误差由 (5.17) 给出．对 (5.17) 进行分部积分，推得当 $b \neq \frac{1}{3}$ 时，$F^{(2)}$ 有精确阶 $p=2$，(5.17) 为它的误差表示式． 令 $a=b=0$，则 (5.18) 和 (5.17) 分别归结成梯形法和它的局部截断误差式 (5.9)．

作为 $F^{(2)}$ 的特殊情形，对于 $b=\frac{1}{3}$ 和 $y(t)$ 为四次连续可微，我们得到公式

$$F^{(3)}: \; y_{n+1} - y_n - \frac{h}{2}[(1+a)y'_{n+1} + (1-a)y'_n]$$

$$+ \frac{h^2}{12}[(1+3a)y''_{n+1} - (1-3a)y''_n] = 0, \tag{5.19}$$

其误差表达式为

$$e_3(t) = -\frac{h^4}{12}\int_0^1 \theta[2\theta^2 - 3(1-a)\theta$$

$$+ (1-3a)]y^{(4)}(t+\theta h)d\theta. \tag{5.20}$$

作为 $F^{(3)}$ 的特殊情形，对于 $a=0$ 和 $y(t)$ 为五次连续可微，得到公式

$$y_{n+1} - y_n - \frac{h}{2}(y'_{n+1} + y'_n) + \frac{h^2}{12}(y''_{n+1} - y''_n) = 0, \tag{5.21}$$

它对应于第二对角线 Padé 近似，精确阶 $p=4$．它的误差项为

$$e_4(t) = \frac{h^5}{24}\int_0^1 \theta^2(\theta-1)^2 y^{(v)}(t+\theta h)d\theta. \tag{5.22}$$

将公式 $F^{(2)}$ 和 $F^{(3)}$ 应用到数试验方程 (5.5)，得到的数值解 y_n 满足递推式

$$y_{n+1} = r^{(\nu)}(q)y_n, \quad \nu = 2,3, \tag{5.23}$$

其中

$$r^{(2)}(q) = [4 + 2(1 - a)q + (b - a)q^2][4 - 2(1 + a)q + (b + a)q^2]^{-1} \tag{5.24}$$

和

$$r^{(3)}(q) = [12 + 6(1 - a)q + (1 - 3a)q^2][12 + 6(1 + a)q + (1 + 3a)q^2]^{-1}. \tag{5.25}$$

类似于公式 $F^{(1)}$，容易证明公式 $F^{(2)}$ 为 A 稳定的充分必要条件是

$$a \geqslant 0, \quad b \geqslant 0. \tag{5.26}$$

对于公式 $F^{(3)}$,有 $b = \dfrac{1}{3}$, A 稳定性的必要充分条件(5.26)缩减成

$$a \geqslant 0. \tag{5.27}$$

对任意给定的实数 q_0，可以找到拟合到阶 $\phi = 0$ 的公式 $F^{(3)}$,记得到的公式为 $F^{(3)}[q_0]$. 这时 $a = a^{(3)}(q_0)$,其中

$$a^{(3)}(q) = \frac{1}{3}[q^2 + 6q + 12 - e^q(q^2 - 6q + 12)] \cdot [e^q(q^2 - 2q) + q^2 + 2q]^{-1}. \tag{5.28}$$

公式 $F^{(2)}$ 含有二个自由参数，利用它们可以在二个地方 q 和 q' 处进行 $\phi = 0$ 阶的指数拟合. 用 $F^{(2)}[q, q']$ 表示在 q 和 q' 处拟合所得到的公式 $F^{(2)}$. 这种拟合就是求参数 a 和 b,使有 $\varepsilon^{(2)}(q) = \varepsilon^{(2)}(q') = 0$,其中

$$\varepsilon^2(q) = r^{(2)}(q) - e^q. \tag{5.29}$$

这样,我们得到确定参数 a 和 b 的二个线性代数方程,它们的形式解为

$$\begin{aligned} a^{(2)}(q, q') &= 2[\tilde{r}(q') - \tilde{r}(q)][q\tilde{r}(q') - q'\tilde{r}(q)]^{-1}, \\ b^{(2)}(q, q') &= 2(q - q')[q\tilde{r}(q') - q'\tilde{r}(q)]^{-1}, \end{aligned} \tag{5.30}$$

式中

$$\tilde{r}(q) = q^2(1 - e^q)[-(2 + q) + (2 - q)e^q]^{-1}. \tag{5.31}$$

如果 (q, q') 是一对实数 (q_1, q_2),或者是一对共轭复数 $(q_0, \bar{q}_0)$,

由 (5.30) 可得到实参数 (a, b). 在后一种情形, (5.30) 可用 $q_0 = \xi + i\eta$ 的实部 ξ 和虚部 η 来表示,

$$
\begin{aligned}
a^{(2)}(\xi, \eta) &= 2[(\xi^2 + \eta^2)(\xi \sin\eta + \eta \sinh\xi) \\
&\quad - 4\xi\eta(\cosh\xi - \cos\eta)][D(\xi, \eta)]^{-1}, \\
b^2(\xi, \eta) &= -2\eta[(\xi^2 + \eta^2 + 4)\cosh\xi + (\xi^2 \\
&\quad + \eta^2 - 4)\cos\eta - 4(\xi \sinh\xi + \eta \sin\eta)][D(\xi, \eta)]^{-1},
\end{aligned}
\tag{5.32}
$$

其中

$$
D(\xi, \eta) = (\xi^2 + \eta^2)[(\xi^2 + \eta^2)\sin\eta - 2\eta(\cosh\xi - \cos\eta)]. \tag{5.33}
$$

在导出 (5.30) 时,我们假定 $q \neq q'$. 但是当 q 固定而 $q' \to q$ 时, 仍能得到合理的表达式, 即在 $a^2(q_1, q_2)$ 和 $b^{(2)}(q_1, q_2)$ 中令 $q_2 \to q_1$, 而在 $a^{(2)}(\xi, \eta)$ 和 $b^2(\xi, \eta)$ 中令 $\eta \to 0$. 在这种极限下, 分别在 q_1 或在 $(\xi, 0)$ 处得到阶 $\phi = 1$ 的拟合.

如果我们已用指数拟合确定了公式 $F^{(\nu)}$, $\nu = 1,2,3$ 的参数, 将这些参数固定,考虑下面二个问题:

(i) 这个公式是否是 A 稳定的.

(ii) 对于 $|q| \ll 1$, 指数拟合公式是否精确. 考虑问题 (ii) 是因为使公式 $F^{(\nu)}$ 为 A 稳定的参数空间中的区域在某些方向上是无界的,即使拟合公式是 A 稳定的,它的参数在绝对值上可任意大. 由于局部截断误差线性地依赖于这些参数, 对于任意给定的 h,这个误差也可能任意的大,所以存在精确性的问题.

为了回答这些问题,我们首先对例 5.1 和例 5.2 中构造的公式定义一个性质,称这个性质为中间性.将公式 $F^{(\nu)}$ 写成统一的形式

$$
y_{n+1} - y_n - h[c_1 y'_{n+1} + c_2 y'_n] + h^2[c_3 y''_{n+1} - c_4 y''_n] = 0. \tag{5.34}
$$

对于 $\nu = 1$, $c_3 = c_4 = 0$. 由向量 $c = (c_1, c_2, c_3, c_4)$ 可用来刻划任意指定的公式. 特别,将与梯形公式 $\left(F^{(1)}, \mu = \dfrac{1}{2}\right)$ 有关的向量 c 记成 $s^{(1)} = \left\{\dfrac{1}{2}, \dfrac{1}{2}, 0, 0\right\}$, 与向后 Euler 公式 $(F^{(1)}, \mu=0)$

有关的向量记成 $b^{(1)}=\{1,0,0,0\}$. 公式 (5.21) 的向量记成 $s^{(2)}=\left\{\frac{1}{2},\frac{1}{2},\frac{1}{12},\frac{1}{12}\right\}$. 将 $a=b=1$ 的 $F^{(2)}$ 所表示的公式的向量记成 $b^{(2)}=\left\{1,0,\frac{1}{2},0\right\}$. 记号 s 和 b 分别表示对称的和向后的意义. b 是一个与向后 Taylor 展开公式有关的向量.

定义 5.3 公式 $F^{(\nu)}$ 称作是中间的,如果有

$$\min(s_j^{(\sigma)}, b_j^{(\sigma)}) \leqslant c_j \leqslant \max(s_j^{(\sigma)}, b_j^{(\sigma)}), j=1,2,3,4,$$

其中对于 $\nu=1$,用 $\sigma=1$,而对于 $\nu=2,3$,用 $\sigma=2$.

对于 $\nu=1$ 和 $j=3,4$. 上面的不等式是平凡满足的. 在这个定义的意义下,所谓中间公式,是指它们位于二个极限公式之间,一个是与对角线 Padé 近似有关的公式,另一个是与在适当次数截断的向后 Taylor 展开有关的公式.

容易证明,下面的条件对于中间性是必要和充分的:

$$F^{(1)}:\ 0\leqslant\mu\leqslant\frac{1}{2}, \tag{5.35}$$

$$F^{(2)}:\ \frac{1}{3}\leqslant a+b\leqslant 2,\ 0\leqslant b-a\leqslant\frac{1}{3}, \tag{5.36}$$

$$F^{(3)}:\ 0\leqslant a\leqslant\frac{1}{3}. \tag{5.37}$$

由 (5.36) 定义的区域是图 5.1 中的矩形. 约束条件 (5.35), (5.36), (5.37) 分别推出例 5.1 和例 5.2 中的 A 稳定性条件. 因此有下面的定理

定理 5.1 形为 $F^{(\nu)}$, $\nu=1,2,3$ 的中间公式均是 A 稳定的.

还容易证明中间公式对于 $|q|\ll 1$ 是精确的, 这就是下面的定理.

定理 5.2 形式为 $F^{(\nu)}$ 的中间公式的局部截断误差对参数的变化按估计式

$$|e_\nu(t)|\leqslant h^{\nu+1}\max_{0<\theta<1}|y^{(\nu+1)}(t+\theta h)|/\nu! \tag{5.38}$$

是一致有界的.

下面的四个定理建立了指数拟合和中间性之间的关系. 本质上, 定理 5.3—定理 5.6 的证明可以这样进行. 首先将指数拟合要求的参数的适当表示式 (5.15), (5.28) 或 (5.30) 代入中间性条件 (5.35), (5.36), (5.37), 或者代入 A 稳定性条件 (5.26), 这些条件将表示成一个变量或二个变量的超越不等式, 然后再证明对于负实轴上的所有可能的拟合点均满足这些不等式. 或者在定理 5.6 的情形,左半复平面的适当子集将满足这些不等式.

定理 5.3 对于任意的 q_0, $-\infty \leqslant q_0 \leqslant 0$, 指数拟合公式 $F^{(1)}[q_0]$ 是中间的. 反过来, 对于任意指定的中间公式 $F^{(1)}$, 可以找到唯一的 q_0, $-\infty \leqslant q_0 \leqslant 0$,使得 $F^{(1)} \equiv F^{(1)}(q_0)$.

定理 5.4 对于任意的 q_0, $-\infty \leqslant q_0 \leqslant 0$, 指数拟合公式 $F^{(3)}[q_0]$ 是中间的.

定理 5.5 对于任意实数对 (q_1, q_2), $-\infty \leqslant q_1, q_2 \leqslant 0$, 指数拟合公式 $F^{(2)}[q_1, q_2]$ 是中间的.

定理 5.6 考虑图 5.2 中表示的区域 $D_2 \subset D_1 \subset \{z \mid \mathrm{Re}\, z < 0\}$.

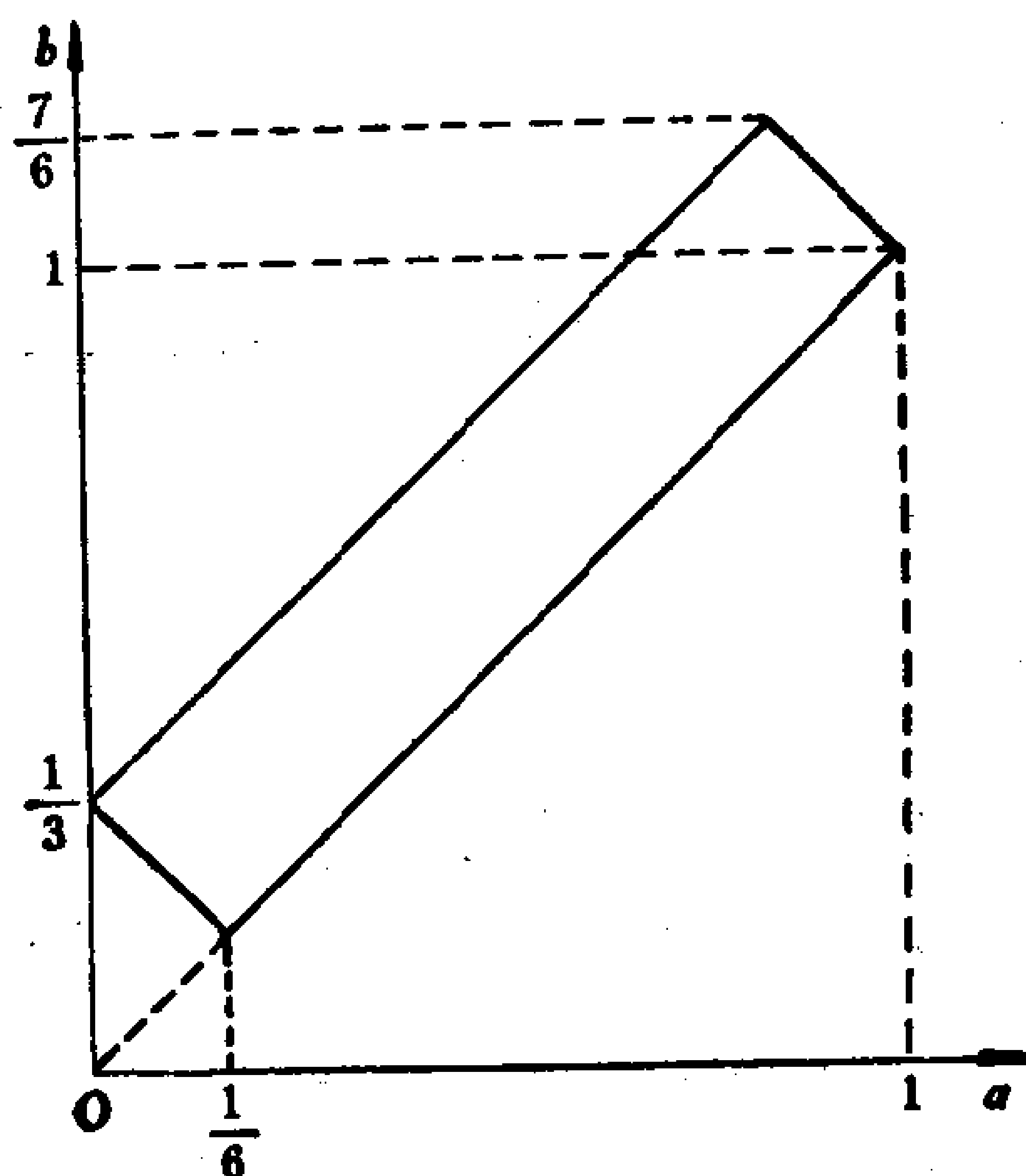


图 5.1 $F^{(2)}$ 的中间性区域

对于 $q_0 \in D_1$，指数拟合公式 $F^{(2)}[q_0, \bar{q}_0]$ 是 A 稳定的. 对于 $q_0 \in D_2$ 指数拟合公式 $F^{(2)}[q_0, \bar{q}_0]$ 是中间的.

当 $\xi \to -\infty$ 时，D_1 的边界渐近地有 $\eta \sim e^{-\xi}$，而 D_2 的边界渐近地有 $\eta \sim \sqrt{2}(-\xi)^{\frac{1}{4}}e^{-\xi/4}$，这里 $q_0 = \xi + i\eta$. 定理 5.6 对于能

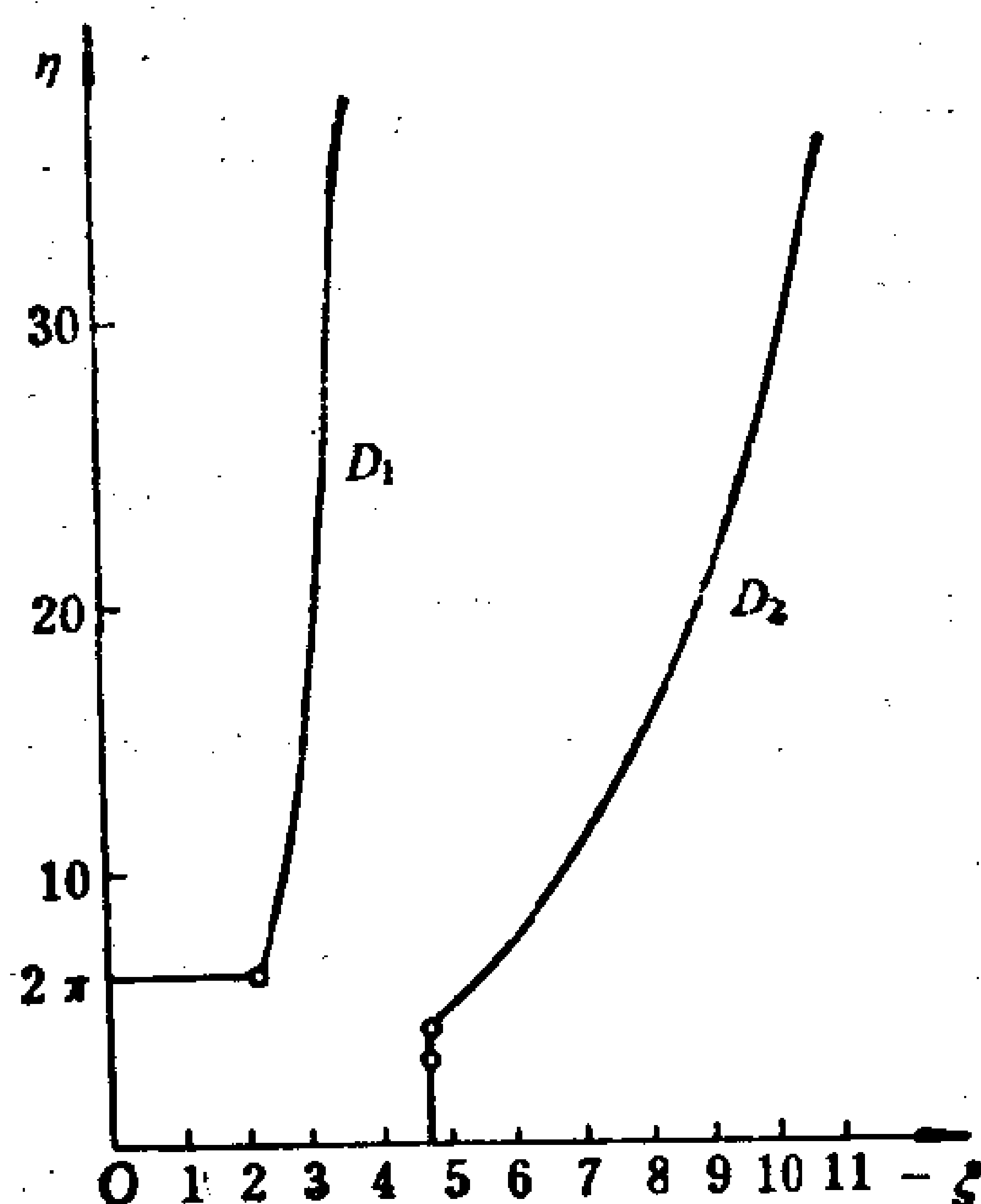


图 5.2 定理 5.6 的区域 D_1 和 D_2

进行指数拟合的复数值 q_0 给出了一些限制. 事实上，自由参数将用来拟合快速衰减的解分量，而不管这些解分量是否是振荡的. 一般说来，这样的 q_0 将有 $|q_0| \gg 1$，因而 q_0 将位于这粗糙的范围内. 定理 5.6 中所包含的条件仅是充分的.

在将指数拟合方法应用到刚性方程组时，必须要有一个选取拟合点的策略. 如果计算了 Jacobi 矩阵 $J = \dfrac{\partial f}{\partial y}$，我们可以确定它的"大的"特征值(例如用幂方法). 再按下面的方式来选取参数：如果存在一个接近于实轴的大特征值的聚点，可以在这个聚点处拟合 $F^{(1)}$ 或 $F^{(3)}$. 如果存在二个接近于实轴的分开的聚点，可以在这些聚点的实平均值 q_1 和 q_2 处拟合 $F^{(2)}$. 类似地，对于具有共轭平均值 q_0 和 $\bar{q}_0$ 的二个分开的共轭复聚点，只要它们满足定理

5.6 中的约束，就可以在这二个点处对 $F^{(2)}$ 进行指数拟合． 根据 Jacobi 矩阵 $J(t)$ 的变化速度，在适当的时间区间上可以重新计算这个矩阵的特征值，并求出自由参数的值． 另外一种选取的方法是参数只选取一次． 特别可以在 $-\infty$ 处拟合 $F^{(1)}$ 或 $F^{(3)}$，或者可以用一个自由参数让 $F^{(2)}$ 在 $-\infty$ 处拟合．在只拟合一次的情形，给定的 q_0 值变化的范围内存在一个在某种意义下为最好的选取．例如，在求 μ 使

$$E(\mu) = \max_{-\infty \leqslant q \leqslant 0} |r^{(1)}(q) - e^q|$$

达极小的意义下，q 的最好的选取为 $q_0 = -8.19$，对应的 μ 值为 $\mu_0 = 0.122$，$E(\mu_0) = 0.139$．作为比较，$\mu=0$ 时，$E(0)=0.204$，而 $\mu = \frac{1}{2}$ 时，$E\left(\frac{1}{2}\right) = 1$．

§2 应用广义 Hermite-Birkhoff 内插的指数拟合多步方法

考虑初值问题

$$y' = f(t, y),\ y(0) = y_0, \tag{5.39}$$

假定有唯一的一次连续可微解 $y(t)$． 用等距格点集 $I_m = \{t_i = ih | i = 0, 1, \cdots, m, mh = T\}$ 上的值 y_i 来逼近 $y(t_i)$． 对于指数拟合，再引进比定义 5.2 更一般的定义

定义 5.4 一个数值方法称作在 λ 处是 p 次指数拟合的，如果对于问题 (5.39) 的形式为 $p(t)e^{\lambda t}$ 的解能精确地求解，其中 $p(t)$ 是任何次数不超过 p 的多项式．即如果用的是精确的初始条件，并且计算是精确的，则方法将得到精确的理论解 $p(t_i)e^{\lambda t_i}$，$i = 0, 1, \cdots, m$．

考虑由集合

$$\left\{t^j e^{\lambda_i t} | i = 0, 1, \cdots, s, j = 0, 1, \cdots, p_i, \sum_{i=0}^{s} (p_i + 1) = r,\right.$$

$\lambda_0=0$, λ_i 是给定的实常数，若 $i\neq k$，则 $\lambda_i\neq\lambda_k\}$

展成的线性空间 Λ_{rs}. 由定义 5.4 推出，若对于解在 Λ_{rs} 中的问题 (5.39)，方法能精确求解时，方法在 λ_i, $i=0, 1, \cdots, s$ 处是 p_i 次指数拟合方法.

这一节利用由广义 Hermite-Birkhoff 内插多项式来构造的形式为

$$y_n=\sum_{i=1}^{k}\alpha_i(h)y_{n-i}+h\sum_{i=0}^{k}\beta_i(h)f(t_{n-i}, y_{n-i}) \tag{5.40}$$

的线性多步方法，使得这方法能精确地积分解在 Λ_{rs} 中的问题. 特别，当 $s=0$ 时，Λ_{r0} 是次数不超过 $r-1$ 的多项式空间，这样构造的方法就是常系数(与 h 无关)的线性多步方法.

令 r, p 是给定的正整数，$h>0$ 是固定的步长. 若方法为隐式，取 $p=k$; 若为显式，取 $p=k-1$. 作 $(p+1)\times 2$ 矩阵 $E=(e_{ij})$，其元素 $e_{ij}=0$ 或 1 ($i=0, 1, \cdots, p$, $j=0, 1$)，并且 $\Sigma_{i,j}e_{ij}=r$，我们称矩阵 E 为作用矩阵. 令 U 是给定的(内插)函数空间，其基为 $\{u_i|u_i: R\rightarrow R,\ i=0, 1, \cdots, r-1\}$. 下面假定空间 U 满足推移不变性，即

$$U|[-kh, 0]=\tau_z U|[z-kh, z] \tag{5.41}$$

对任何 $z\in R$ 成立. 这里 τ_z 是推移算子: $\tau_z u(t)=u(t-z)$. (5.41) 的意义是这样的: 对每个 $u(t)\in U|[-kh, 0]$ 存在一个 $v(t)\in U|[z-kh, z]$，使 $u(t)=\tau_z v(t)=v(t-z)$ 对每个 $t\in[-kh, 0]$ 均成立；反过来也成立. 由作用矩阵 E 和内插空间可以定义下面的 Hermite-Birkhoff 内插问题: 寻找 $u(t)\in U$，使对每个 $e_{ij}=1$ 的对 (i, j) 成立

$$u^{(j)}(t_{n-k+i})=y^{(j)}(t_{n-k+i}), \tag{5.42}$$

这里 $y^{(0)}(t_{n-i})$ 表示 y_{n-i}，而 $y^{(1)}(t_{n-i})$ 表示 $f_{n-i}=f(t_{n-i}, y_{n-i})$. 取作用矩阵 E 和内插空间 U 使上述问题有唯一解 $u(t)\in U$. 例如，取 $U=\Lambda_{rs}$，而 E 有形式

$$E = \begin{bmatrix} 0 & & 1 \\ \vdots & & \vdots \\ 0 & & 1 \\ 1 & & \\ \vdots & \boxed{\diagup} & \\ 1 & & \\ 0 & & 1 \\ \vdots & & \vdots \\ 0 & & 1 \end{bmatrix}. \tag{5.43}$$

E 的第一列至少具有一个 1，而位置 $\boxed{\diagup}$ 上的元素可以为零也可以为 1. 空间 Λ_{rs} 显然是推移不变的. 这样构造的 E 和 U 满足上述问题的要求. 下面假定 E 具有 (5.43) 的形式.

用内插问题的解 $u(t)$ 来构造微分方程 (5.39) 的解在格点 t_n 处的值的近似，定义 $y_n = u(t_n)$. 由推移不变性，每个 $u(t) \in U$ 有表示式

$$u(t_{n-i}) = \sum_{j=0}^{r-1} c_j u_j(-ih), \tag{5.44}$$

于是由条件 (5.42) 导出方程

$$Ac = g, \tag{5.45}$$

其中 $c = (c_0, c_1, \cdots, c_{r-1})^T$. A 是 $r \times r$ 矩阵，由行向量 $(u_0^{(j)}(-ih), u_1^{(j)}(-ih), \cdots, u_{r-1}^{(j)}(-ih))$ 所组成，其中 (i, j) 使 $e_{k-i,j} = 1$. g 是 r 维向量，具有 $e_{ij} = 1$ 的分量 $y^{(j)}(t_{n-k+i})$. 方程 (5.45) 给出

$$c = A^{-1}g \tag{5.46}$$

和

$$y_n = u(t_n) = \sum_{j=0}^{r-1} c_j u_j(0) = v^T A^{-1} g, \tag{5.47}$$

其中记 $v = (u_0(0), u_1(0), \cdots, u_{r-1}(0))^T$. 向量 $v^T A^{-1}$ 的元仅依赖于 h，因此公式 (5.47) 具有 (5.40) 的形式，这只要将对应于

$e_{ii}=0$ 的 $y^{(i)}(t_{n-i})$ 的系数看成为零就可以看出.

由 Λ_{rs} 的推移不变性和插值问题解的唯一性推出这样构造的方法对解在 Λ_{rs} 中的每个问题将能精确积分.

例 5.3 取 $U=\Lambda_{r0}$ 来构造，得到常系数线性多步方法. 对解为不超过 $r-1$ 次的多项式的问题，这些方法能精确求解. 因此有阶 $r-1$. 不同的作用矩阵将导出不同的方法. 取

$$E=\begin{bmatrix}1 & 1\\ 0 & 1\end{bmatrix} \tag{5.48}$$

和 $U=\Lambda_{30}$，这时 $u_0(t)=1$, $u_1(t)=t$, $u_2(t)=t^2$. 由 (5.48)，方程 (5.45) 有形式

$$\begin{pmatrix}u_0(-h) & u_1(-h) & u_2(-h)\\ u_0'(0) & u_1'(0) & u_2'(0)\\ u_0'(-h) & u_1'(-h) & u_2'(-h)\end{pmatrix}\begin{pmatrix}c_0\\ c_1\\ c_2\end{pmatrix}=\begin{pmatrix}y(-h)\\ y'(0)\\ y'(-h)\end{pmatrix}.$$

现在矩阵 A 为

$$A=\begin{pmatrix}1 & -h & h^2\\ 0 & 1 & 0\\ 0 & 1 & -2h\end{pmatrix},$$

求得 c_0, c_1, c_2 为

$$c_0=y(-h)+\frac{h}{2}(y'(0)+y'(-h)),$$

$$c_1=y'(0),$$

$$c_2=[y'(-h)-y'(0)]/(-2h),$$

因而

$$y(t_n)=y(t_{n-1})+\frac{h}{2}(y'(t_n)+y'(t_{n-1})),$$

这就是熟知的梯形法.

例 5.4 取 $U=\Lambda_{21}$，基函数为 $1, t, e^{\lambda t}$，则得到指数拟合公式，E 仍取 (5.48)，这时矩阵 A 为

$$A=\begin{pmatrix}1 & -h & e^{-\lambda h}\\ 0 & 1 & \lambda\\ 0 & 1 & \lambda e^{-\lambda h}\end{pmatrix},$$

求得 c_0, c_1, c_2 为

$$c_0 = y(-h) + \frac{-(\lambda h + 1)e^{-\lambda h}}{\lambda(1 - e^{-\lambda h})} y'(0)$$

$$+ \frac{\lambda h + e^{-\lambda h}}{\lambda(1 - e^{-\lambda h})} y'(-h),$$

$$c_1 = \frac{-e^{-\lambda h}}{1 - e^{-\lambda h}} y'(0) + \frac{1}{1 - e^{-\lambda h}} y'(-h),$$

$$c_2 = \frac{1}{\lambda(1 - e^{-\lambda h})} y'(0) - \frac{1}{\lambda(1 - e^{-\lambda h})} y'(-h),$$

得到的积分公式为

$$y(t_n) = c_0 u_0(0) + c_1 u_1(0) + c_2 u_2(0)$$

$$= y(t_{n-1}) + \frac{1 - (\lambda h + 1)e^{-\lambda h}}{\lambda(1 - e^{-\lambda h})} y'(t_n)$$

$$+ \frac{\lambda h - 1 + e^{-\lambda h}}{\lambda(1 - e^{-\lambda h})} y'(t_{n-1}).$$

下面我们来研究指数拟合多步方法 (5.40) 的渐近性质. 定理 5.7 表示,若给出了适当的起始程序,并且右函数 $f(t, y)$ 是充分光滑的(即 f 对 t 连续,而对 y 是 Lipschitz 连续),则这些渐近性质将化成对应多项式方法的性质.

定理 5.7 应用任何 Λ_{rs} 和作用矩阵 E 所构造的指数拟合方法为收敛的充分条件是用 Λ_{r0} 和同样的 E 构造的方法是收敛的,并且方法的阶至少是 $r - 1$.

证明 由方法的构造,系数 $\alpha_i(h)$ 和 $\beta_i(h)$ 均是以基函数 $(h^j e^{\lambda_i h})$ 的具不为零的分母的有理表达式,因此对 h 是无限多次可微的. 这推出方法至少是 $r - 1$ 阶的 ($r - 1 = \dim\Lambda_{rs} - 1$, Λ_{rs} 是方法的零子空间). 象对于 Λ_{r0} 的收敛方法必须有 $r \geqslant 2$ 一样,指数拟合方法的阶至少为 1,这推出相容性. 为了证明零稳定性,将 (5.40) 写成

$$y_n = \sum_{i=1}^{k} \alpha_i y_{n-i} + h \left[\sum_{i=1}^{k} a_i(h) y_{n-i} \right.$$

$$+ \sum_{i=0}^{k} \beta_i(h) f(t_{n-i}, y_{n-i}) \Big], \tag{5.49}$$

其中 $\alpha_i(h) = \alpha_i + h a_i(h)$. 如果用 Λ_{r0} 代替 Λ_{rs}, 并且用同样的 E,我们将得到同样的常数 α_i. 这是因为 Λ_{rs} 具有形式为

$$v_i = \sum_{j=i}^{\infty} c_j t^j$$

的基 $\{v_i\}_{i=0}^{r-1}$. 由于 $\alpha_i(h)$ 和 $\beta_i(h)$ 的光滑性推出对小的 h, $a_i(h)$ 和 $\beta_i(h)$ 的有界性. 因此方法 (5.40) 与多项式方法一样,系数 α_i 满足根条件将推出收敛性.

现在来讨论方法的稳定区域的渐近性质. 考虑单个试验问题

$$y' = \mu y, \ y(0) = 0.$$

定义多项式

$$\begin{aligned} \rho(\zeta, h) &= \sum_{i=0}^{k} \alpha_i(h) \zeta^{k-i}, \\ \sigma(\zeta, h) &= \sum_{i=0}^{k} \beta_i(h) \zeta^{k-i}, \\ \alpha_0(h) &\equiv -1. \end{aligned} \tag{5.50}$$

定义 5.5 方法 (5.40) 在 μh 平面上的区域中是稳定的,如果方程

$$\rho(\zeta, h) + \mu h \sigma(\zeta, h) = 0 \tag{5.51}$$

的根按模 $\leqslant 1$,而且模为 1 的根是单根.

为了说明方便,先研究一个拟合参数的情形,并且假定在 λ 处是零次拟合的. 当 $\lambda \to 0$ 时,方法趋向于对应的常系数方法. 因此对于小的 λ 值, 稳定区域接近于对应的常系数方法的区域. 有兴趣的问题是当 $\lambda \to -\infty$ 时, 这些区域是如何变化的, 特别要考察是否扩大方法的稳定区域.

为简单起见,选取

$$E=\begin{pmatrix}1 & 1\\ 1 & 1\\ \vdots & \vdots\\ 1 & 1\\ 0 & 1\end{pmatrix},$$

用 $u_i(i=0,1,\cdots,r-1)$ 表示基函数，并取 $u_{r-1}(t)=e^{\lambda t}$，$u_i(t)=t^i$，$i=0,1,\cdots,r-2$. 这样构造内插的矩阵 A 为

$$A=\begin{pmatrix}u_0(-h) & \cdots & u_{r-2}(-h) & e^{-\lambda h}\\ u_0(-2h) & \cdots & u_{r-2}(-2h) & e^{-2\lambda h}\\ \vdots & & \vdots & \vdots\\ u_0(-kh) & \cdots & u_{r-2}(-kh) & e^{-k\lambda h}\\ u_0'(0) & \cdots & u_{r-2}'(0) & \lambda\\ u_0'(-h) & \cdots & u_{r-2}'(-h) & \lambda e^{-\lambda h}\\ \vdots & & \vdots & \vdots\\ u_0'(-kh) & \cdots & u_{r-2}'(-kh) & \lambda e^{-k\lambda h}\end{pmatrix}.$$

由方法的构造可以看到，在构造过程中 A^{-1} 是最重要的因素. 给定一个固定的 h，为了考察 $\lambda\to-\infty$ 时的情形，可用代数余子式来构造 A^{-1}. 按 A 的最后一列元素展开，

$$\det A=\sum_{i=1}^{k}a_i e^{-i\lambda h}+\lambda\sum_{i=0}^{k}b_i e^{-i\lambda h},$$

其中系数 a_i 和 b_i 不依赖于 λ. 特别有 $b_k\neq 0$，因为它对应于由 $\{u_0,u_1,\cdots,u_{r-2}\}$ 展成的空间 $\widetilde{U}$ 及矩阵

$$\widetilde{E}=\begin{pmatrix}1 & 0\\ 1 & 1\\ \vdots & \vdots\\ 1 & 1\\ 0 & 1\end{pmatrix}$$

构造的内插中的矩阵 $\widetilde{A}$ 的行列式(可能符号不同). 记

$$A^{-1}=\frac{1}{\det A}\operatorname{Adj}A=\left(\frac{A_1 \vdots A_2 \vdots \cdots \vdots A_r}{B}\right),$$

其中 $A_i, i=1,2,\cdots,r$ 是列向量，B 是行向量. 它们的元素是代数余子式除以 $\det A$. B 的元素有形式 $w/\det A$，w 不依赖于 λ，并且当 $\lambda\to-\infty$ 时，$w/\det A\to 0$. 列向量 A_i，$i=1,\cdots,r-1$ 的元素有形式

$$f(\lambda)=\frac{\sum_{i=1}^{k} c_i e^{-i\lambda h}+\lambda\sum_{i=0}^{k} d_i e^{-i\lambda h}}{\sum_{i=1}^{k} a_i e^{-i\lambda h}+\lambda\sum_{i=0}^{k} b_i e^{-i\lambda h}},$$

其中余子式是按最后一列的元素展开的. 由于 $b_k\neq 0$，当 $\lambda\to-\infty$ 时，$f(\lambda)\to d_k/b_k<\infty$. 这里 d_k 是应用 $\tilde{E}$ 和 $\tilde{U}$ 得到的低维问题中的对应余子式. 最后，向量 A_r 的元素有形式

$$q(\lambda)=\frac{\sum_{i=1}^{k} c_i e^{-i\lambda h}+\lambda\sum_{i=0}^{k-1} d_i e^{-i\lambda h}}{\sum_{i=1}^{k} a_i e^{-i\lambda h}+\lambda\sum_{i=0}^{k} b_i e^{-i\lambda h}}.$$

当 $\lambda\to-\infty$ 时，$q(\lambda)\to 0$ $\left(因为 \frac{c_k}{\lambda b_k}\to 0\right)$，于是推出

$$\lim_{\lambda\to-\infty} A^{-1}=\left[\begin{array}{c|c}\tilde{A}^{-1} & 0\\ \hline 0 & 0\end{array}\right],$$

其中 $\tilde{A}^{-1}$ 是 $(r-1)\times(r-1)$ 矩阵，不依赖于 λ. 向量 $A^{-1}g$ 的极限是一个向量，其最后元素为零，并且没有一个元素是依赖于 g 的最后一个元 f_{n-k} 的.

上面实际上已指出，$\tilde{A}^{-1}$ 是用 $\tilde{E}$ 和 $\tilde{U}$ 进行内插构造的矩阵 $\tilde{A}$ 的逆. 我们注意，当 $\lambda\to-\infty$ 时，对应于 (5.47) 的极限结果为

$$y_n=\sum_{i=0}^{r-1} c_i u_i(0),$$

恰好是由 $\tilde{E}$ 和 $\tilde{U}$ 的低维内插得到的，而不需用极限过程. 因此可以看到，当 $\lambda\to-\infty$ 时，方程 (5.51) 的根趋向于

$$\tilde{\rho}(\zeta, h) + \mu h \tilde{\sigma}(\zeta, h) = 0$$

的根，其中 $\tilde{\rho}$ 和 $\tilde{\sigma}$ 是低维方法的特征多项式.

即使 E 上有多个零也得到类似的结果.

下面假如 l 个 $(l > 1)$ 基函数依赖于 λ，而 $\tilde{U}$ 表示不依赖于 λ 的其余的基函数 $\{u_i\}_{i=0}^{r-1-l}$ 所展成的空间. 对应地，记 $\tilde{E}$ 为按下列规则将 E 中 l 个 1 换成 0 所得到的作用矩阵. 从含 1 的最上面的行开始，并从右边到左边用 0 代替这些 1，直到已处理完 l 个 1 为止. 令这样充满 0 的行数为 $k - \tilde{k}$. 将由 $\tilde{U}$ 和 $\tilde{E}$ 构造的低维方法的特征多项式表成 $\tilde{\rho}$ 和 $\tilde{\sigma}$. 现在如果在 $\tilde{E}$ 的第一列中仍至少包含一个 1，则上面的推理导出结果：当 $\lambda \to -\infty$ 时，用 U 和 E 构造的方法等价于用 $\tilde{U}$ 和 $\tilde{E}$ 构造的方法，并且方程 (5.51) 的根趋向于

$$\zeta^{k-\tilde{k}}[\tilde{\rho}(\zeta, h) + \mu h \tilde{\sigma}(\zeta, h)] = 0$$

的根. 如果在 $\tilde{E}$ 的第一列中只含零，则上面的推理不成立. 在这种情形中，行列式 b_k 为零，需要进行另外的处理. 现在假定出现这种情况. 令 $e_{k-m,0}$ 是从第一列中除去的最后一个 1，于是有 $\tilde{e}_{i,0} = 0, i = 0, 1, \cdots, p$ 和 $\tilde{e}_{i1} = 0, i = 0, 1, \cdots, k - \tilde{k} - 1$ 和 $\tilde{e}_{i1} = 1, i = k - \tilde{k}, k - \tilde{k} + 1, \cdots, p$，所以 $\tilde{E}$ 有 $\tilde{p} + 1 = p + 1 - k + \tilde{k}$ 个非零行. 对于隐式方法 $k = p$，而对于显式方法 $k = p + 1$. 类似地，对于低维的情形 $\tilde{k} = \tilde{p}$ 或 $\tilde{k} = \tilde{p} + 1$. 通过与前面类似的分析，我们可以看到当 $\lambda \to -\infty$ 时，$\beta_i \to 0(i = k, k - 1, \cdots, \tilde{k} + 1)$，$\alpha_m \to 1$，和如果 $m < k$，则 $\alpha_i \to 0(i = k, k - 1, \cdots, m + 1)$ 和 $|\beta_i| \to \infty(i = \tilde{k}, \tilde{k} - 1, \cdots, i_0)$，其中对于隐式 $i_0 = 0$，而对显式 $i_0 = 1$. 对于显式方法的情形，(5.51) 中 ζ^k 的系数是常数. 当 $\mu h \neq 0$ 时，若由 $\lambda \to -\infty$ 推得 $|\beta_i| \to \infty(i = 1, \cdots, \tilde{k})$，则至少有一个根按模无限增大. 在隐式方法的情形，当 $\mu h \neq 0$ 及 $\lambda \to -\infty$ 时，ζ^k 的系数的模有 $|h\mu\beta_0 - 1| \to \infty$，用这个系数来除进行规格化，我们考虑表达式 $\dfrac{\mu h \beta_i}{\mu h \beta_0 - 1}$ 的极限. 通过与前面类似的计算，可以看到，这些极限与 μh 的值无关，等于

$(-1)^i\binom{\tilde{k}}{i}$, $i=0,1,\cdots,\tilde{k}$. 所以 (5.51) 的根收敛于

$$\zeta^{k-\tilde{k}}(\zeta-1)^{\tilde{k}}=0$$

的根。

将上面的分析总结成下面的定理。

定理 5.8 如果 $\tilde{E}$ 的第一列中至少含一个 1，则当 $\lambda\to-\infty$ 时，方程 (5.51) 的根趋向于

$$\zeta^{k-\tilde{k}}[\tilde{\rho}(\zeta,h)+\mu h\tilde{\sigma}(\zeta,h)]=0$$

的根；否则，当 $\mu h\neq 0$ 时，如果方法是显式的，则至少有一个根按模趋向于无穷大；如果它是隐式的，则方程 (5.51) 的根趋向于

$$\zeta^{k-\tilde{k}}(\zeta-1)^{\tilde{k}}=0$$

的根。

由这个定理推出，只要 $\tilde{E}$ 的第一列至少含有一个 1 和 $\tilde{\sigma}\not\equiv 0$，极限方法的稳定区域是与应用 $\tilde{U}$ 和 $\tilde{E}$ 的低维方法的稳定区域相同的。但是如果 $\tilde{\sigma}\equiv 0$，极限方法的根就不依赖于 μh，我们的分析就得不出关于稳定区域的任何性质。例如对指数拟合的 Euler 方法就属于这种情形：对任何复 μh，(5.51) 的唯一根趋向于 1，而稳定区域的极限是左半平面。

由定理还可以推得下面的结论：如果我们希望在左半平面上有任意大的稳定区域，U 不能包含次数高于 2 的多项式。这是因为不存在阶大于 2 的常系数的 A 稳定的多步方法。

例 5.5 考虑一些 Adams 型的简单方法。选

$$E=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix},$$

我们得到形式为

$$y_n=y_{n-1}+h[\beta_0(h)f_n+\beta_1(h)f_{n-1}]$$

的隐式单步方法。在表 5.1 中列出了应用不同的内插空间的基和得到的相应的系数函数。这些方法的稳定区域是圆。如果 $\beta_1(h)\neq\beta_0(h)$，则圆心在 $(1/[\beta_0(h)-\beta_1(h)],0)$ 处；如果 $\beta_1(h)=\beta_0(h)$，则稳定区域是左半平面。对不同的 λ 方法 II 的稳定区域在图 5.3 中画出。当 $\lambda\to-\infty$ 时，稳定区域趋向于应用

$$E=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \text{ 和 } \{1, t\}$$

的向后 Euler 方法的稳定区域，这也在图 5.3 中表出。

对于显式方法，我们选

$$E=\begin{pmatrix}0 & 1\\ 1 & 1\end{pmatrix}.$$

表 5.1 隐式单步方法的系数函数，情形 I—IV

情　形	I	II	III	IV
U 的基	$\{1, t, t^2\}$	$\{1, t, e^{\lambda t}\}$	$\{1, e^{\lambda t}, te^{\lambda t}\}$	$\{1, e^{\lambda t}, e^{\mu t}\}$
$\beta_0(h)$	$\frac{1}{2}$	$\frac{1+\lambda h-e^{\lambda h}}{\lambda h(1-e^{\lambda h})}$	$\frac{e^{-\lambda h}+\lambda h-1}{\lambda^2h^2}$	$\frac{\lambda-\mu-\lambda e^{\mu h}+\mu e^{\lambda h}}{\lambda\mu h(e^{\lambda h}-e^{\mu h})}$
$\beta_1(h)$	$\frac{1}{2}$	$\frac{e^{\lambda h}-1-\lambda h e^{\lambda h}}{\lambda h(1-e^{\lambda h})}$	$\frac{e^{\lambda h}-\lambda h-1}{\lambda^2h^2}$	$\frac{\mu e^{\mu h}-\lambda e^{\lambda h}+(\lambda-\mu)e^{(\lambda+\mu)h}}{\lambda\mu h(e^{\lambda h}-e^{\mu h})}$

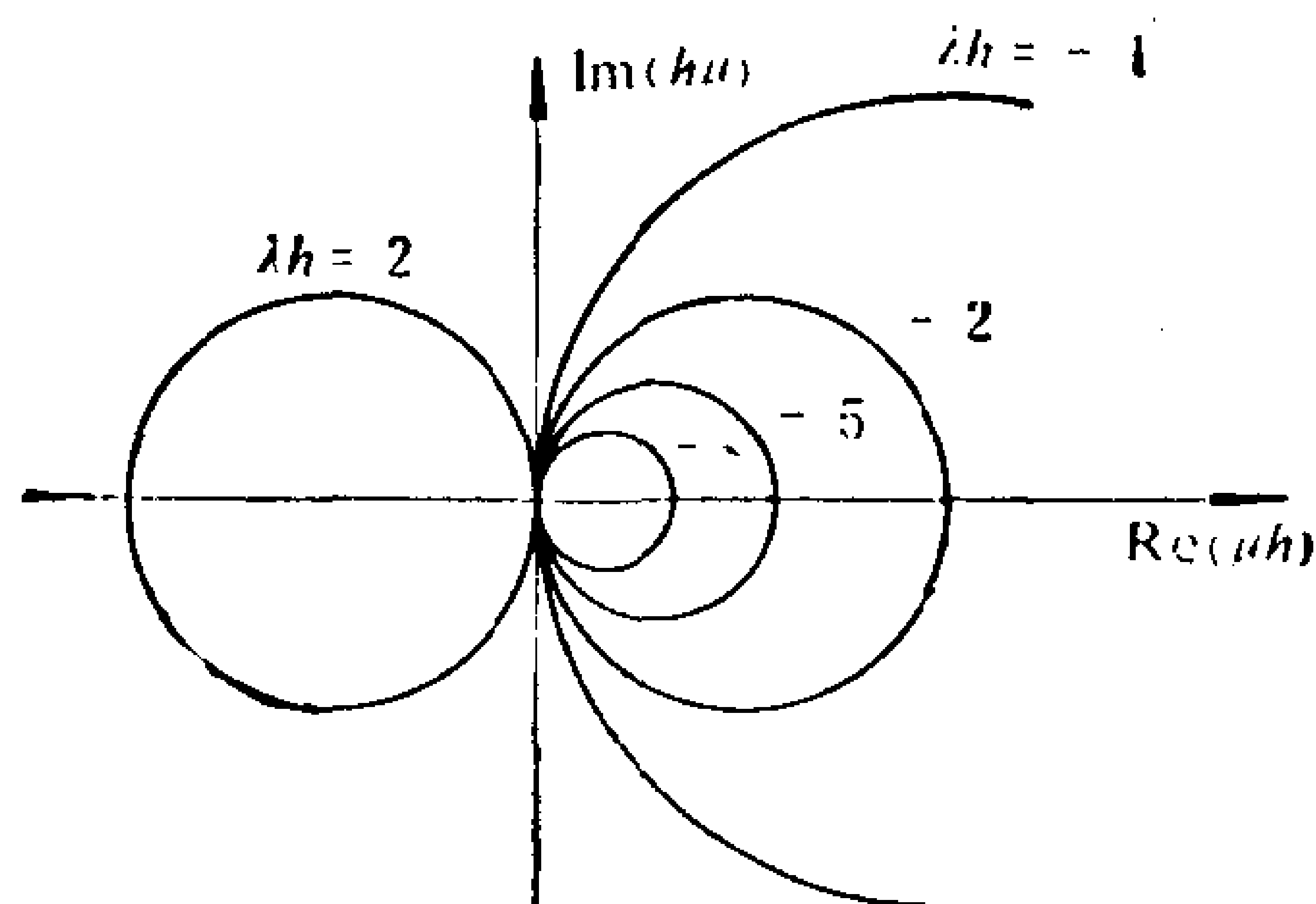


图 5.3　方法 II 的稳定区域．对于 $\lambda h<0$，方法在圆外和圆上是稳定的，对 $\lambda h>0$，方法在圆内和圆上是稳定的．

(通常将 E 中为零的行略去)，在 λ 处，次数为 0 和 1 的指数拟合和在 λ, μ 处次数为 0 的指数拟合，均列在表 5.2 中，这些方法具有形式

$$y_n=y_{n-1}+h[\beta_1(h)f_{n-1}+\beta_2(h)f_{n-2}].$$

方法 V—VII 的稳定区域在图 5.4 中给出．在 VI 的情形，当 $\lambda\to-\infty$ 时，极限区域是用 $E=(1, 1)$ 和 $\{1, t\}$ 的 Euler 方法的区域，而在 VII 的情形，是左半平面．

表 5.2 显式二步方法的系数函数，情形 V—VIII

情 形	V	VI	VII	VIII
U 的基	$\{1, t, t^2\}$	$\{1, t, e^{\lambda t}\}$	$\{1, e^{\lambda t}, te^{\lambda t}\}$	$\{1, e^{\lambda t}, e^{\mu t}\}$
$\beta_1(h)$	$\frac{3}{2}$	$\frac{e^{\lambda h}-1-\lambda he^{-\lambda h}}{\lambda h(1-e^{-\lambda h})}$	$\frac{e^{-\lambda h}(1-\lambda h)+2\lambda h-1}{\lambda^2h^2e^{-\lambda h}}$	$\frac{\lambda e^{-\lambda h}-\mu e^{-\mu h}+\mu e^{(\lambda-\mu)h}-\lambda e^{(\mu-\lambda)h}}{\lambda\mu h(e^{-\mu h}-e^{-\lambda h})}$
$\beta_2(h)$	$-\frac{1}{2}$	$\frac{1+\lambda h-e^{\lambda h}}{\lambda h(1-e^{-\lambda h})}$	$\frac{e^{\lambda h}-\lambda he^{\lambda h}-1}{\lambda^2h^2e^{-\lambda h}}$	$\frac{\mu-\lambda-\mu e^{\lambda h}+\lambda e^{\mu h}}{\lambda\mu h(e^{-\mu h}-e^{-\lambda h})}$

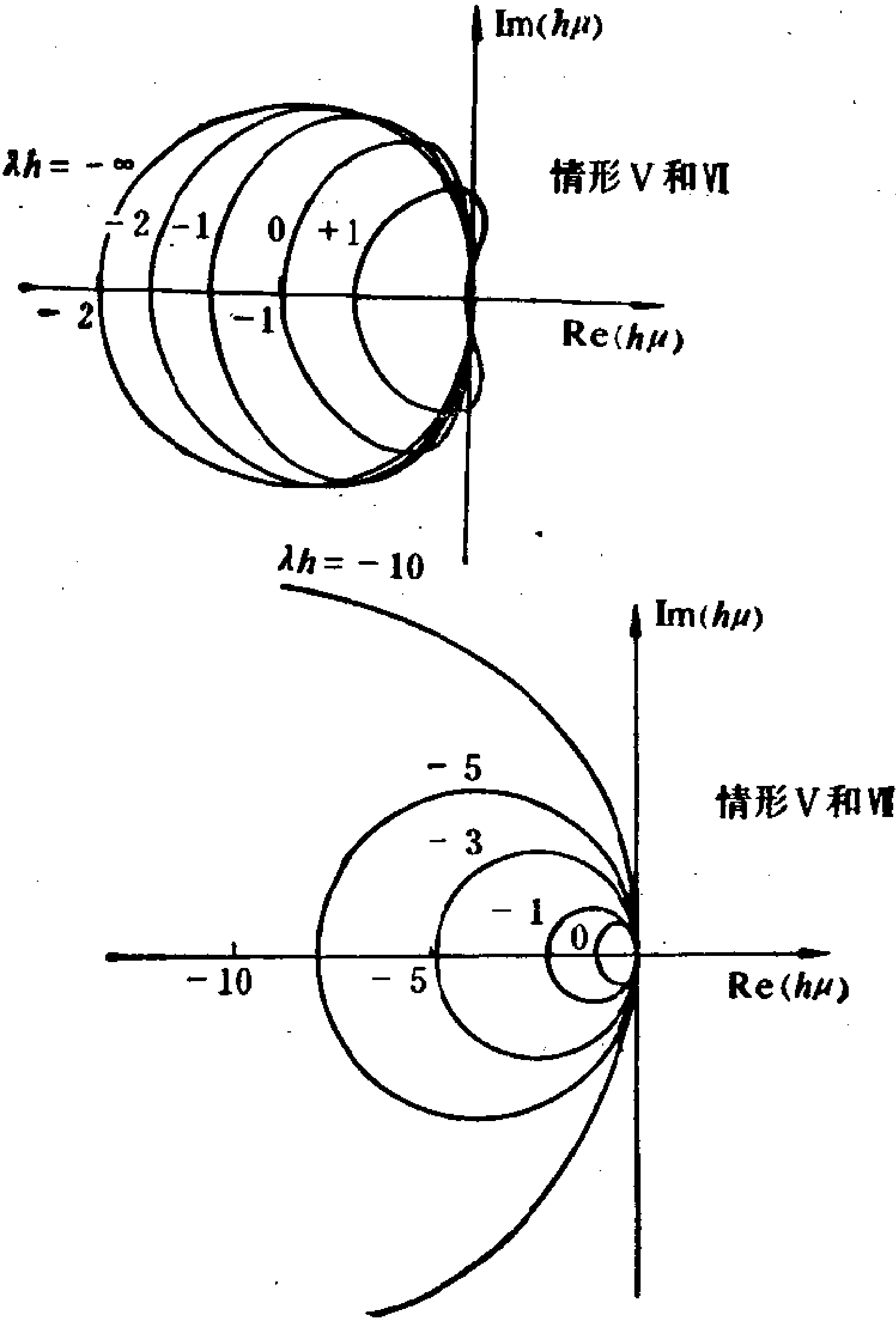


图 5.4 情形 V，VI，VII 的稳定区域，在闭图的内部和图上方法是稳定的，$\lambda h=0$ 为V的情形.

对于隐式二步方法，取

$$E=\begin{pmatrix}0 & 1\\ 1 & 1\\ 0 & 1\end{pmatrix},$$

在 λ 处次数为 0, 1, 2 的拟合公式在表 5.3 中列出，这些公式具有形式

$$y_n=y_{n-1}+h[\beta_0(h)f_n+\beta_1(h)f_{n-1}+\beta_2(h)f_{n-2}].$$

情形 X 和 XI 的稳定区域在图 5.5 中给出，$\lambda\to-\infty$ 的极限区域包含左半平面，即梯形公式(情形 X) 和向后 Euler 公式(情形 XI)。

表 5.3 隐式二步方法的系数函数

情 形	IX	X
U 的基	$\{1, t, t^2, t^3\}$	$\{1, t, t^2, e^{\lambda t}\}$
$\beta_0(h)$	$\frac{5}{12}$	$\frac{2+3\lambda h-\lambda h e^{-\lambda h}-2e^{\lambda h}}{4\lambda h[1-\cosh(\lambda h)]}$
$\beta_1(h)$	$\frac{8}{12}$	$\frac{-4-4\lambda h\cosh(\lambda h)-2\lambda h\sinh(\lambda h)+4e^{\lambda h}}{4\lambda h[1-\cosh(\lambda h)]}$
$\beta_2(h)$	$-\frac{1}{12}$	$\frac{2+2\lambda h-\lambda h(1-e^{\lambda h})-2e^{\lambda h}}{4\lambda h[1-\cosh(\lambda h)]}$

情 形	XI	XII
U 的基	$\{1, t, e^{\lambda t}, te^{\lambda t}\}$	$\{1, e^{\lambda t}, te^{\lambda t}, t^2e^{\lambda t}\}$
$\beta_0(h)$	$\frac{(e^{\lambda h}-2)(1-\lambda h)+e^{-\lambda h}(1-\lambda h-\lambda^2h^2)}{2\lambda^2h^2[1-\cosh(\lambda h)]}$	$\{e^{-\lambda h}(2-\lambda h)-2+3\lambda h-2\lambda^2h^2\}/\{-2\lambda^3h^3\}$
$\beta_1(h)$	$\frac{2\lambda h(\lambda h-1)+1-e^{2\lambda h}+e^{\lambda h}(1+\lambda h)-e^{-\lambda h}(1-\lambda h)}{2\lambda^2h^2[1-\cosh(\lambda h)]}$	$\frac{4(1-\lambda h)e^{\lambda h}+2\lambda^2h^2-4}{-2\lambda^3h^3}$
$\beta_2(h)$	$\frac{e^{2\lambda h}-e^{\lambda h}(2+\lambda^2h^2)+1}{2\lambda^2h^2[1-\cosh(\lambda h)]}$	$\frac{e^{\lambda h}(2+\lambda h)+e^{2\lambda h}(\lambda h-2)}{-2\lambda^3h^3}$

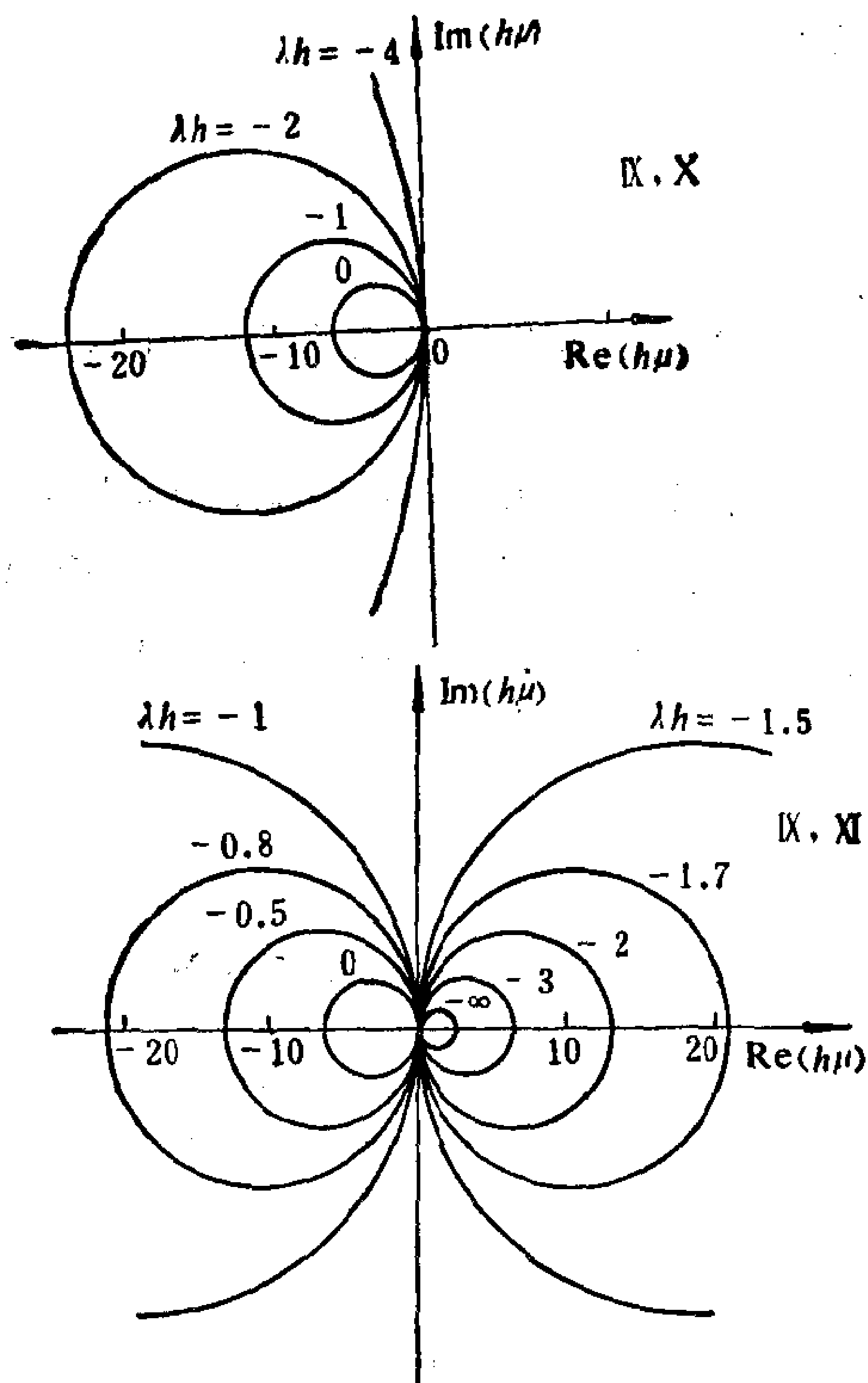


图 5.5 情形 IX, X, XI 的稳定区域，在左半平面的图内和图上是稳定的，而在右半平面的图上和图外是稳定的，$\lambda h=0$ 为 IX 的情形.

§3 矩阵多步方法的指数拟合

这一节利用指数拟合的思想将一般的 Adams 型公式作一些改变，使其能适合于求解刚性方程组.

考虑积分 N 阶常微分方程组的线性多步公式

$$\sum_{i=-(k-1)}^{1} \alpha_{i+k-1}(Q)y_{n+i}-h\sum_{i=-(k-1)}^{1} \beta_{i+k-1}(Q)y'_{n+i}=0, \tag{5.52}$$

其中系数$\alpha_i(Q)$和$\beta_i(Q)$均是N阶对角矩阵，依赖于N阶参数对角线矩阵Q. 在下面记号F_k表示显式("预估")k步公式($\beta_k=0$),记号F_k^*表示隐式("校正")k步公式($\beta_k\neq 0$). $\bar{F}_k$和$\bar{F}_k^*$分别表示F_k, F_k^*的数量分量(行). 我们将推导公式F_k和F_k^*,使对任意固定的Q,当将公式应用到

$$y' = -Dy + \phi(t, y) \tag{5.53}$$

时,得到的解是精确的,只要其中$Q = hD$和$\phi(t, y(t))$是独立变量t的次数$\leqslant k-1$或$\leqslant k$的任意多项式. 每个行$\bar{F}_k$或$\bar{F}_k^*$只依赖于$q = hd$,其中d和q分别为D和Q的对应的对角线元素.

任意的N阶方程组

$$y' = f(t, y) \tag{5.54}$$

均可以改写成形式(5.53),只要令

$$\phi = Dy + f(t, y). \tag{5.55}$$

当然，在某种意义下ϕ相对于f或Dy很小时，公式(5.52)将是特别适用的. 即使公式(5.52)是由$\tilde{Q} = h\tilde{D}$, $\tilde{D} \neq D$构成的,而不是用Q,则下面推导的公式也是有意义的,也可以与传统的数值方法相比拟. 事实上,对于$q = 0$, $\bar{F}_k$或$\bar{F}_k^*$即是通常的显式或隐式的k步 Adams 公式.类似地,乘上适当的因子,对于$q = +\infty$, $\bar{F}_k^*$是k步向后差分公式，而对于任意的q, $0 \leqslant q \leqslant +\infty$, $\bar{F}_k^*$即是§1中考虑的指数拟合公式. 这里要注意，这一节中的q与§1中的q的意义差一个符号.

公式(5.52)对于数值求解刚性方程组是有用的. 因为当在大的q处拟合时，$\bar{F}_k^*$将具有一些强的固定步长的稳定性质，可以用来控制对应的"快速过渡过程",当对于小的$|q|$值拟合时,对于任意k, $\bar{F}_k$和$\bar{F}_k^*$是稳定的，并且它们将非常适合于处理非刚性分量.

§3.1 积分公式的推导

令$\{t_n\}$为 $t_n = nh$, $n = 0, 1, \cdots, h > 0$. 方程(5.53)可以转换成积分方程

$$y(t_{n+1}) = e^{-hD}y(t_n) + \int_{t_n}^{t_{n+1}} e^{-(t_{n+1}-t)D}\phi(t, y(t))dt, \quad (5.56)$$

其中 t_n 是固定的，t_{n+1} 看成是独立变量，$h = t_{n+1} - t_n$. 利用 (5.56)，可以构造一类线性多步公式. 为此，用 t 的多项式来近似 $\phi(t, y(t))$，然后将这个积分算出. 如果这个多项式是通过点 $\phi_{n-i} = \phi(t_{n-i}, y_{n-i})$, $i = 0, 1, \cdots, k-1$, $k \geq 1$ 的 Lagrange 插值多项式，则构造了显式的或预估型式的 k 步公式 $\bar{F}_k$. 如果 Lagrange 插值通过点 ϕ_{n-i+1}, $i = 0, 1, \cdots, k$，则得到隐式的或校正的 k 步公式 $\bar{F}_k^*$.

为了导出 $\bar{F}_k$，应用 Newton 向后差分公式，得到通过点 $\phi_{n-i}(i = 0, 1, \cdots, k-1)$ 的多项式 $\hat{\phi}(t)$. 如果在 (5.56) 中用 $\hat{\phi}$ 代替 ϕ，得到 F_k 的差分形式方程

$$y_{n+1} - e^{-Q}y_n - h\sum_{i=0}^{k-1} G_i \nabla^i \phi_n = 0, \quad k \geq 1, \quad (5.57)$$

其中 ∇^i 表示第 i 阶向后差分，由 $Q = hD$ 和

$$G_i = (-1)^i \int_0^1 e^{-(1-\xi)Q} \binom{-\xi}{i} d\xi, \quad i = 0, 1, \cdots \quad (5.58)$$

计算得到

$$G_i = \sum_{j=0}^{i} \phi_{i,j} Q^{-(j+1)} - e^{-Q} \sum_{j=0}^{i} \xi_{i,j} Q^{-(j+1)}, \quad i \geq 0, \quad (5.59)$$

其中

$$\phi_{i,j} = \frac{j!}{i!} \gamma_{i,j}, \quad i \geq 0, \ 0 \leq j \leq i, \quad (5.60)$$

$$\xi_{i,j} = \frac{1}{i!} \sum_{l=j}^{i} \gamma_{i,l} l!/(l-j)!, \quad 0 \leq j \leq i. \quad (5.61)$$

而 $\gamma_{i,j}$, $i \geq 0$, $0 \leq j \leq i$ 由公式

$$\binom{-\xi}{i} \equiv \frac{(-1)^i}{i!} \sum_{j=0}^{i} \gamma_{i,j} \zeta^j, \quad i \geq 0 \quad (5.62)$$

来确定，$\zeta = 1 - \xi$. 显然，$\gamma_{0,0} = \gamma_{1,0} = 1$ 和 $\gamma_{1,1} = -1$，容易验证 $\gamma_{i,j}$ 满足递推关系

$$\left.\begin{aligned} &\gamma_{i+1,0} = (i+1)\gamma_{i,0}, \\ &\gamma_{i+1,i+1} = -\gamma_{i,i}, \\ &\gamma_{i+1,j} = (i+1)\gamma_{i,j} - \gamma_{i,j-1},\ 1 \leqslant j \leqslant i. \end{aligned}\right\} i \geqslant 1 \tag{5.63}$$

若在 (5.62) 中令 $\zeta = 1$,得到

$$\sum_{j=0}^{i} \gamma_{i,j} = 0,\ i \geqslant 1. \tag{5.64}$$

由此推得

$$\xi_{i,0} = 0,\ i > 1.$$

为了将 F_k 转换成通常的形式,我们作代换

$$\nabla^i \phi_n = \sum_{j=0}^{i} (-1)^j \binom{i}{j} \phi_{n-j},\ i = 0, 1, \cdots, \tag{5.65}$$

这就产生公式

$$y_{n+1} - e^{-Q} y_n - h \sum_{i=0}^{k-1} B_{k-1,i} \phi_{n-i} = 0,\ k \geqslant 1, \tag{5.66}$$

其中

$$\begin{aligned} B_{i,j} &= (-1)^j \sum_{l=j}^{i} \binom{l}{j} G_l \\ &= \sum_{l=0}^{i} \mu_{i,j,l} Q^{-(l+1)} - e^{-Q} \sum_{l=0}^{i} \nu_{i,j,l} Q^{-(l+1)}, \\ &\qquad i \geqslant 0,\ 0 \leqslant j \leqslant i, \end{aligned} \tag{5.67}$$

和

$$\begin{Bmatrix} \mu \\ \nu \end{Bmatrix}_{i,j,l} = (-1)^j \sum_{r=\max(j,l)}^{i} \binom{r}{j} \begin{Bmatrix} \phi \\ \xi \end{Bmatrix}_{r,l}, \quad 0 \leqslant j \leqslant i,\ 0 \leqslant l \leqslant i, \tag{5.68}$$

这表示 μ 和 ν 分别与 ϕ 和 ξ 有关.

可以对 F_k^* 进行类似于上面的推导. 与隐式公式有关的量我们将用"*"来标出. F_k^* 的差分形式为

$$y_{n+1} - e^{-Q} y_n - h \sum_{i=0}^{k} G_i^* \nabla^i \phi_{n+1} = 0, \tag{5.69}$$

其中

$$G_i^* = (-1)^i \int_{-1}^{0} e^{\xi Q} \binom{-\xi}{i} d\xi$$

$$= \sum_{j=0}^{i} \phi_{i,j}^* Q^{-(j+1)} - e^{-Q} \sum_{j=0}^{i} \xi_{i,j}^* Q^{-(j+1)}, \ i = 0, 1, \cdots, \tag{5.70}$$

这里

$$\left.\begin{aligned} \phi_{i,j}^* &= (-1)^j \frac{j!}{i!} \gamma_{i,j}^* \\ \xi_{i,j}^* &= \frac{1}{i!} \sum_{l=j}^{i} (-1)^l \gamma_{i,l}^* \frac{l!}{(l-j)!} \end{aligned}\right\} 0 \leqslant j \leqslant i, \tag{5.71}$$

量 $\gamma_{i,j}^*$ 由关系式

$$\binom{-\xi}{i} = \frac{(-1)^i}{i!} \sum_{j=0}^{i} \gamma_{i,j}^* \xi^j, \ i \geqslant 0 \tag{5.72}$$

来确定．容易证明，对于所有的 i, j，有 $\gamma_{i,j}^* = \xi_{i,j}$．按定义，对于 $i \geqslant 2$，有 $\xi_{i,0}^* = 0$．

公式 (5.69) 等价于通常的形式

$$y_{n+1} - e^{-Q} y_n - h \sum_{i=0}^{k} B_{k,i}^* \phi_{n+1-i} = 0, \ k \geqslant 1, \tag{5.73}$$

其中

$$B_{i,j}^* = (-1)^j \sum_{l=j}^{i} \binom{l}{j} G_l^*, \ i \geqslant 0, \ 0 \leqslant j \leqslant i, \tag{5.74}$$

$B_{i,j}^*$ 和 G_l^* 之间的关系式 (5.74) 与 $B_{i,j}$ 和 G_l 之间的关系式是相同的．因此，将关系式 (5.67) 和 (5.68) 中的量加上星号后得到的关系式仍成立．

上面构造的公式是传统的 Adams 公式的推广．特别在 $q \to 0$ 的情形，$\bar{F}_k$ 和 $\bar{F}_k^*$ 分别趋向于显式的或隐式的 k 步 Adams 公式．事实上，在 § 3.3 中将证明：如果将 q 看成是独立的常参数，$\bar{F}_k^*$ 的局部截断误差是 $q g_{k+1}^* h^{k+1} y_n^{(k+1)} + O(h^{k+2})$．于是对于 $q = 0$，$\bar{F}_k^*$ 的精确阶为 $p = k + 1$，另外，对于 $q = 0$，按 (5.53)，有 $\phi = y'$．因此 $\bar{F}_k^*$ 只含 y 的前二个项，它是 Adams 型的．由于唯

一性，这个公式确实是熟知的 k 步隐式 Adams 公式. 类似的讨论说明，对于 $q=0$，公式 $\bar{F}_k$ 是阶 $p=k$ 的显式 k 步 Adams 公式.

现在考虑 $q\to+\infty$ 的情形. 如果我们略去对于 q 的负幂(指数上)为小的项，得到 $g_0^*=\phi_{0,0}^* q^{-1}=q^{-1}$ 和 $g_i^*=\phi_{i,1}^* q^{-2}+O(q^{-3})$，$i\geqslant 1$. 于是如果我们令

$$\theta_{k,0}^*=\sum_{i=1}^{k}\phi_{i,1}^*,$$

$$\theta_{k,j}^*=(-1)^j\sum_{l=j}^{k}\binom{l}{j}\phi_{l,1}^*,\ 1\leqslant j\leqslant k. \tag{5.75}$$

并且用 $b_{k,j}^*$ 表示 $B_{k,j}^*$ 的一般的对角线分量，有

$$\begin{aligned} b_{k,0}^* &= q^{-1}+\theta_{k,0}^* q^{-2}+O(q^{-3}),\\ b_{k,j}^* &= \theta_{k,j}^* q^{-2}+O(q^{-3}),\ 1\leqslant j\leqslant k, \end{aligned} \tag{5.76}$$

因此，由于 $h\phi_{n+1-j}=hy'_{n+1-j}+qy_{n+1-j}$. 我们得到

$$\begin{aligned} b_{k,0}^* h\phi_{n+1} &= y_{n+1}+\theta_{k,0}^* q^{-1}y_{n+1}+hq^{-1}y'_{n+1}+O(q^{-2}),\\ b_{k,j}^* h\phi_{n+1-j} &= \theta_{k,j}^* q^{-1}y_{n+1-j}+O(q^{-2}),\ 1\leqslant j\leqslant k. \end{aligned} \tag{5.77}$$

用 q 乘公式 $\bar{F}_k^*$，再取 $q\to+\infty$ 的极限，这个公式变成

$$\sum_{j=0}^{k}\alpha_{k-j}y_{n+1-j}-hy'_{n+1}=0, \tag{5.78}$$

其中系数 α_i, $i=0,1,\cdots,k$ 依赖于 k. 公式 (5.78) 是向后微分型的，即它只含最前的一个 y' 项. 对于任何 $q\neq 0$，由上面给出的公式 $\bar{F}_k^*$ 的截断误差是 $O(h^{k+1})$，即 $\bar{F}_k^*$ 的阶为 $p=k$. 再由唯一性，公式 (5.78) 等同于变 $\beta_k=1$ 后的通常的 k 步向后微分公式.

§3.2 稳定性分析

在对公式 $F_k^*(F_k)$ 进行 A 稳定性分析时，可以按第二章，检验充分性条件 N_1 和 N_2. 通过变换 $w=w(z)=(z+1)/(z-1)$，并令

$$\begin{aligned} r(z) &= (z-1)^k\rho(w(z)),\\ s(z) &= (z-1)^k\sigma(w(z)). \end{aligned}$$

条件 N_1 等价于条件 N_1': $s(z)$ 的根 s_i 有 $\mathrm{Re}\,s_i<0$, $i=1,2,\cdots$,

k. 条件 N_1' 可用 Routh-Hurwitz 定理来验证. 记 $s(z)=a_0z^k+b_0z^{k-1}+\cdots$, 并用 $a_0, b_0, c_0, \cdots$ 表示 Routh 表第一列的元素, 于是当且仅当 $a_0, b_0, c_0, \cdots$ 均为非零的, 并且具有相同的符号时, N_1' 成立

下面对试验方程 $y'=\lambda y$, $\lambda=$ 常数, 讨论固定 h 的稳定性时, 应用记号 $q'=\lambda h$, 以便区分拟合点 $q=dh$.

由 § 1 的公式 (5.8) 和 (5.15), 对于任意的 q, $0\leqslant q\leqslant+\infty$, 公式(这里的 q 与 § 1 的 q 差一个符号)

$$
\begin{aligned}
F_1^*:\ & y_{n+1}-y_n \\
& -h\{[q^{-1}-(e^q-1)^{-1}]y_n'+[(1-q^{-1}) \\
& +(e^q-1)^{-1}]y_{n+1}'\}=0
\end{aligned}
\tag{5.79}
$$

是 A 稳定的.

Liniger [78] 中提出一个阶 $p=2$ 的指数拟合二步公式, 它等价于公式

$$
\begin{aligned}
F_2^*:\ & [(1-2q^{-1})+e^{-q}(1+2q^{-1})]y_{n-1}+4[(-1+q^{-1}) \\
& -e^{-q}q^{-1}]y_n+[(3-2q^{-1})+e^{-q}(-1+2q^{-1})]y_{n+1} \\
& -h\{[(-q^{-1}+2q^{-2})-e^{-q}(q^{-1}+2q^{-2})]y_{n-1}' \\
& +[(4q^{-1}-4q^{-2})-e^{-q}(2-4q^{-2})]y_n'+[(2-3q^{-1} \\
& +2q^{-2})-e^{-q}(-q^{-1}+2q^{-2})]y_{n+1}'\}=0.
\end{aligned}
\tag{5.80}
$$

当 $2<q\leqslant+\infty$ 时, 这个公式是 A 稳定的. 可以这样来证明. 用 g_i^* 表示 G_i^* 的一般的对角线元素. 我们得到 $a_0=g_0^*=[(1-e^{-q})/q]>0$, $q>0$. 令 $\hat{b}_0(q)=q^2b_0/2=q-(1-e^{-q})$, 其中 $b_0=2(g_0^*+g_1^*)$. 于是 $\hat{b}_0'(q)>0$, $q>0$ 和 $\hat{b}_0(0)=0$. 由这推得 $b_0(q)>0$, $q>0$. 最后令 $\hat{c}_0=q^3c_0=(2-q)[(2-q)-(2+q)e^{-q}]$, 其中 $c_0=a_1=g_0^*+2g_1^*+4g_2^*$. 我们有 $[(2-q)-(2+q)e^{-q}]<2-q\leqslant 0$, $q\geqslant 2$. 因此, $\hat{c}_0>0$, $q>2$. 这推出 $c_0>0$, $q>2$. 即对于 $q>2$, 条件 N_1' 满足. 条件 N_2 要求有

$$
P(\xi,q)=2r(q)(\xi-1)^2\geqslant 0,\ -1\leqslant\xi\leqslant 1, \tag{5.81}
$$

其中

$$
r(q)=q^{-5}r_1(q)r_2(q),
$$

$$r_1(q) = \frac{1}{2}[(2-q)-(2+q)e^{-q}],$$

$$r_2(q) = (-2+3q-q^2)+(2-q)e^{-q}.$$

如上所述，对于 $q>2$，$r_1(q)<0$. 但 $r_2(q)<-q^2+3q-2=(q-2)(1-q)<0$, $q>2$. 因此 $r(q)>0$, $q>0$ 和对所有 ξ, $-1\leqslant\xi\leqslant 1$, $q>2$ 有 $P(\xi,q)\geqslant 0$. 这推出了所要的结果.

公式 $\bar{F}_3^*$ 有形式

$$\begin{aligned}
&[(11-12q^{-1}+6q^{-2})-e^{-q}(2-6q^{-1}+6q^{-2})]y_{n+1}\\
&\quad+[(-18+30q^{-1}-18q^{-2})-e^{-q}(3+12q^{-1}\\
&\quad-18q^{-2})]y_n+[(9-24q^{-1}+18q^{-2})-e^{-q}(-6\\
&\quad-6q^{-1}+18q^{-2})]y_{n-1}+[(-2+6q^{-1}-6q^{-2})\\
&\quad-e^{-q}(1-6q^{-2})]y_{n-2}-h\{[(6-11q^{-1}+12q^{-2}\\
&\quad-6q^{-3})-e^{-q}(-2q^{-1}+6q^{-2}-6q^{-3})]y'_{n+1}+[(18q^{-1}\\
&\quad-30q^{-2}+18q^{-3})-e^{-q}(6-3q^{-1}-12q^{-2}\\
&\quad+18q^{-3})]y'_n+[(-9q^{-1}+24q^{-2}-18q^{-3})\\
&\quad-e^{-q}(6q^{-1}+6q^{-2}-18q^{-3})]y'_{n-1}+[(2q^{-1}-6q^{-2}\\
&\quad+6q^{-3})-e^{-q}(-q^{-1}+6q^{-3})]y'_{n-2}\}=0.
\end{aligned}\tag{5.82}$$

对于 $q\geqslant 5$，条件 N'_1 满足. 为了证明这个结论，只需证明 a_0, b_0, $c_0=a_1-(a_0b_1/b_0)$ 和 $d_0=b_0$ 均不为零，并且对于 $q\geqslant 5$ 具有相同的符号. 首先，对于 $q>0$ 有

$$a_0=6(1-e^{-q})>0,$$

对于 $q\geqslant 2$ 有

$$\begin{aligned}
b_0&=(18-12q^{-1})+(-6+12q^{-1})e^{-q}\\
&\geqslant(18-12q^{-1})+(-6+12q^{-1})=12>0.
\end{aligned}$$

记

$$d_0=2q^{-3}[p_1(q)+p_2(q)e^{-q}],$$

可以看到，对于 $q\geqslant 2$ 有 $p'_2(q)>0$. 再由 $p_2(2)>0$ 推出对于 $q\geqslant 2$, $p_2(q)>0$. 类似地，由当 $q\geqslant 4$ 时，$p'_1(q)>0$ 和 $p_1(5)>0$ 推出对于所有 $q\geqslant 5$ 有 $p_1(q)>0$. 因而 $d_0>0$. 最后，在 $b_0>0$ 的条件下（对于 $q\geqslant 2$ 它是成立的），$c_0>0$ 的充分必要条

件是

$$a_1b_0 - a_0b_1 = q^2N(q)/48 > 0,$$

其中

$$N(q) = (6q^2 - 13q + 9) + (8q - 12)e^{-q} - (q - 3)e^{-2q}$$

对于 $q > 3$, 有 $8q - 12 > 0$ 和 $q - 3 > 0$. 因此,对所有 $q \geqslant 5$

$$N(q) > (6q^2 - 13q + 9) - (q - 3)$$
$$= 6q^2 - 14q + 12 > 0$$

这就推出当 $q \geqslant 5$ 时条件 N_1' 是成立的.

对于 $q \geqslant 5$,公式 F_3^* 还具有一些固定步长 h 的稳定性质. 这可以对有关条件的数值计算来得到. 特别可得到对于 $\alpha \leqslant \alpha_0(q)$ 时, F_3^* 是 $A(\alpha)$ 稳定的. 其中最大角 $\alpha_0(q)$ 从 $q = 5$ 的近似为 84.4°(1.4731 弧度)增加到 $q = +\infty$ 的近似为 86°. 这后一个方法为向后微分公式.

对于 $q = +\infty$, 公式 F_k^* 是向后微分公式. 对所有 k,在∞处都是稳定的. 由于这些公式的系数连续地依赖 q, 则对于任意给定的 k 和充分大的 q 值, $F_k^*(q)$ 必定在∞处的稳定. 类似地,对于 $q = 0$, 公式 F_k 和 F_k^* 均是 Adams 型公式(在 $q \to 0$ 的极限下), 它们是稳定的. 于是对于任何 k 和充分小的非零值 q, $\bar{F}_k$ 和 $\bar{F}_k^*$ 是稳定的.

当然, 对于大的 q 值在∞处为稳定的 $F_k^*(q)$ 可能对于这个 q 值不同时是稳定的. 关于这一点的最明显的例子是对于 $k \geqslant 7$ 的向后微分公式. 它们是不稳定的. 类似地,与小的 q 值对应的近似 Adams 型公式 $F_k(q)$ 和 $F_k^*(q)$ 在∞处均不是稳定的. 这表示对于 q 值, 公式的稳定性质缺少一致性. 这个缺点可以用下面的方式来克服. 在矩阵积分格式中, 用对应于在大的 q 处拟合的行分量公式来处理刚性分量的解, 于是这些公式恰好具有控制刚性分量所需要的稳定性质. 另一方面, 慢变分量(非刚性分量)的解用在小的 q 值处拟合的近似 Adams 型公式处理.

公式 F_3^* 具有上述的一致性, 对于所有的 q 值, $0 \leqslant q \leqslant +\infty$, 它均是稳定的. 为了证明这一点,令 $\rho(w, q)$ 是与 F_3^* 的 α 系数

有关的三次多项式,并令 $\hat{\rho}(w, q)=\rho(w, q)/(w-1)$. 由相容性,对所有 q, $\rho(w, q)$ 具有线性因子 $(w-1)$. 因此, $\hat{\rho}(w,q)$ 是二次多项式. 另外,令

$$\begin{aligned}\hat{p}(z, q) &= (z-1)^2\hat{\rho}((z+1)/(z-1)) \\ &= 2c_0(q)+3qc_1(q)z+3q^2c_2(q)z^2,\end{aligned}$$

其中

$$\begin{aligned}c_0(q) &= (6-9q+5q^2)+(-6+3q+q^2)e^{-q}, \\ c_1(q) &= (-2+3q)+(2-q)e^{-q}, \\ c_2(q) &= 1-e^{-q}.\end{aligned}$$

现在将 Routh 准则应用到 $\hat{p}(z, q)$ 上. 首先, 对于 $q>0$ 有 $c_2(q)>0$. 其次,对于 $0<q\leqslant 3$, 有 $c_1'(q)\geqslant q>0$ 和 $c_1(0)=0$. 因此对于 $0<q\leqslant 3$, $c_1(q)>0$. 对于 $q\geqslant 3$, $c_1(q)\geqslant 2q>0$. 这样对于所有 q, $0<q\leqslant+\infty$, $c_1(q)>0$. 最后,对于 $q\geqslant q_1=(-3+\sqrt{33})/2$, 有 $-6+3q+q^2\geqslant 0$ 和 $c_0(q)\geqslant 6-9q+5q^2>0$; 对于 $0<q\leqslant q_1$ 有

$$\begin{aligned}c_0(q) &\geqslant (6-9q+5q^2)+(-6+3q+q^2)[1-q \\ &\quad +(q^2/2)]=(q^3+q^4)/2>0,\end{aligned}$$

这完成了 F_3^* 对所有 q 为稳定的证明.

§3.3 局部截断误差分析

考虑与公式 F_{p+1} 有关的线性微分差分算子

$$L_{p+1}=\sum_{j=-p}^{1}\alpha_{j+p}(Q)E^j-h\sum_{j=-p}^{0}\beta_{j+p}(Q)E^j\frac{d}{dt}$$

其中 E 是推移算子. L_{p+2} 的阶比 L_{p+1} 的阶高 1, 所以 L_{p+1} 的局部截断误差的主部是与 $L_{p+1}-L_{p+2}$ 相同的. 我们得到

$$(L_{p+1}-L_{p+2})y(t)=hG_{p+1}\nabla^{p+1}[y'(t)+Dy(t)], \quad (5.83)$$

其中用 $y'+Dy$ 代替 ϕ. 对于 (5.53) 的光滑解 $y(t)$,有

$$\begin{aligned}(L_{p+1}-L_{p+2})y(t) &= QG_{p+1}\nabla^{p+1}y(t)+hG_{p+1}\nabla^{p+1}y'(t) \\ &= h^{p+1}QG_{p+1}y^{(p+1)}(t)+O(h^{p+2}).\end{aligned} \quad (5.84)$$

于是我们推得

$$L_{p+1}y(t)=h^{p+1}QG_{p+1}y^{(p+1)}(t)+O(h^{p+2}),\qquad(5.85)$$

当$Q\neq 0$时，F_{p+1} 的精确阶为 p. 当$Q=0$时，精确阶为 $p+1$. 这个结果在显式 Adams 公式中是常见的. 按照(5.85)，在一般的情形，L_{p+1} 的主要误差项的系数是 $c_{p+1}=QG_{p+1}$.

公式 F_p^* 的误差分析与对 F_{p+1} 的分析类似. 设

$$L_p^*=\sum_{j=-(p-1)}^{1}\alpha_{j+p-1}^*(Q)E^j-h\sum_{j=-(p-1)}^{1}\beta_{j+p-1}^*(Q)E^j\frac{d}{dt}\qquad(5.86)$$

是与公式 F_p^* 有关的线性微分差分算子. 我们可以得到

$$L_p^*y(t)=h^{p+1}QG_{p+1}^*y^{(p+1)}(t)+O(h^{p+2}).\qquad(5.87)$$

因此，对于 $Q\neq 0$, F_p^* 的精确阶为 p. 对于 $Q=0$，得到 p 步隐式 Adams 公式的熟知结果，即 F_p^* 的精确阶是 $p+1$. L_p^* 的主误差项的系数是 $c_{p+1}^*=QG_{p+1}^*$.

下面应用 Milne 方法(见 Henrici 的书[62]第 255 页)写出 F_{p+1} 和 F_p^* 的局部截断误差的近似表达式. 令 $L_{p+1}\bar{y}_n^0=0$ 确定的预估值 $\bar{y}_{n+1}^0$. 假定在 $t_{n-p}, t_{n-p+1},\cdots,t_n$ 处 L_{p+1} 作用在精确解上，我们找到

$$\alpha_{p+1}[\bar{y}_{n+1}^0-y(t_{n+1})]\simeq -h^{p+1}c_{p+1}y^{(p+1)}(t_n),\qquad(5.88)$$

其中$\simeq$表示含误差 $O(h^{p+2})$ 的等号. 类似地，由 $L_p^*y_n^1=0$ 确定校正值 y_{n+1}^1，其中 y'_{n+1} 取 $\bar{\Phi}_{n+1}^0-D\bar{y}_{n+1}^0$, $\bar{\Phi}_{n+1}^0=\phi(t_{n+1},\bar{y}_{n+1}^0)$. 于是

$$\alpha_p^*[y_{n+1}^1-y(t_{n+1})]\simeq -h^{p+1}c_{p+1}^*y^{(p+1)}(t_n).\qquad(5.89)$$

由(5.88)和(5.89)消去 $y(t_{n+1})$，求出 $y^{(p+1)}(t_n)$，得到

$$y^{(p+1)}(t_n)=h^{-(p+1)}[\alpha_{p+1}^{-1}c_{p+1}-\alpha_p^{*-1}c_{p+1}^*]^{-1}(y_{n+1}^1-\bar{y}_{n+1}^0)+O(h),\qquad(5.90)$$

α_{p+1} 是恒等矩阵，而

$$\alpha_p^*=I-QB_{p,0}^*.$$

由(5.88)和(5.90)，求得 F_{p+1} 的局部截断误差的一个估计式为

$$\bar{y}_{n+1}^{0} - y(t_{n+1}) \approx \alpha_p^* \hat{G}_{p+1}[\hat{G}_{p+1}^* - \alpha_p^* \hat{G}_{p+1}]^{-1}(y_{n+1}^{1} - \bar{y}_{n+1}^{0}), \tag{5.91}$$

其中

$$\hat{G}_{p+1} = (I + Q)G_{p+1},$$
$$\hat{G}_{p+1}^* = (I + Q)G_{p+1}^*.$$

(5.91) 式是将由 (5.88),(5.90) 得到的式子分别用 $\hat{G}_{p+1}$, $\hat{G}_{p+1}^*$ 代替 G_{p+1}, G_{p+1}^* 得到的. 这样代替得到的公式使用起来要方便一点。因为当 $q \to +\infty$ 时, $\hat{G}(q) \sim qG(q)$ 和 $\hat{G}^*(q) \sim qG^*(q)$ 趋向于有限的非零极限。另一方面,对于 $q \to 0$ 时, $\hat{G} \sim G$, $\hat{G}^* \sim G^*$ 也有合理的极限.

按照完全类似的方式, 对于 F_p^* 的局部截断误差, 得到估计式

$$y_{n+1}^{1} - y(t_{n+1}) \approx \hat{G}_{p+1}^*[\hat{G}_{p+1}^* - \alpha_p^* \hat{G}_{p+1}]^{-1}(y_{n+1}^{1} - \bar{y}_{n+1}^{0}). \tag{5.92}$$

误差估计 (5.91) 和 (5.92) 可以用作步长控制,还可以用它们来对预估值和校正值进行修正.

令修正后的值在 (5.91) 和 (5.92) 中起 $y(t_{n+1})$ 的作用,并将这些关系式中的近似等号用严格的等号代替, 解出修正值. 然后再将预估的修正值公式中的 $(\bar{y}^0 - y^1)$ 的下标向后推移 1. 这样我们得到下面的单步预估修正—校正修正算法

$$\bar{y}_{n+1}^{0} = e^{-Q}y_n + h \sum_{i=-(p-1)}^{1} B_{p,i+p-1}\phi_{n-(p-1)-i}, \tag{5.93}$$

$$\bar{y}_{n+1} = \bar{y}_{n+1}^{0} + \alpha_p^* \hat{G}_{p+1}[\hat{G}_{p+1}^* - \alpha_p^* \hat{G}_{p+1}]^{-1}(\bar{y}_n^{0} - y_n^{1}), \tag{5.94}$$

$$y_{n+1}^{1} = e^{-Q}y_n + h \left[B_{p,0}^*\phi_{n+1} + \sum_{i=-(p-2)}^{1} B_{p,p-1+i}^*\phi_{n+1-(p-1)-i}\right], \tag{5.95}$$

$$y_{n+1} = y_{n+1}^{1} + \hat{G}_{p+1}^*[\hat{G}_{p+1}^* - \alpha_p^* \hat{G}_{p+1}]^{-1}(\bar{y}_{n+1}^{0} - y_{n+1}^{1}), \tag{5.96}$$

其中 $\bar{\phi}=\phi(t,\bar{y})$, $\hat{G}_{p+1}=(I+Q)G_{p+1}$ 和 $\hat{G}^*_{p+1}=(I+Q)G^*_{p+1}$. 在 $Q\to 0$ 的极限情形,(5.93)—(5.96) 变成通常的 Adams 型的修正预估校正型算法.

有时将 (5.93)—(5.96) 中的后二个关系式改成下面的迭代式

$$y^i_{n+1}=e^{-Q}y_n+h\left[B^*_{p,0}\phi^{i-1}_{n+1}\right.$$

$$\left.+\sum_{j=-(p-2)}^{1}B^*_{p,p-1-j}\phi_{n+1-(p-1)-j}\right],\quad i=1,2,\cdots,$$

其中 $\phi^i=\phi(t,y^i)$ 和 $\phi^0=\phi$. 当相邻两个 i 所对应的 y 值的差小于指定的界时，迭代终止. 称这种迭代格式为预估修正迭代校正程序.

§3.4 矩阵 Q 的选取

上面讨论时,矩阵 Q 是固定的. 在实际计算时. 矩阵 Q 可以动态选取. 例如对于方程组 (5.54) 可取 Q 的对角线元素 q_i 为

$$q_i=-h\frac{f^i(t_n,y_n)-f^i(t_n,y^1_n,\cdots,y^i_{n-1},\cdots,y^N_n)}{y^i_n-y^i_{n-1}}. \tag{5.97}$$

这时计算程序中需要计算矩阵 e^{-Q} 的子程序. 当公式 (5.66) 中的 Q 由 (5.97) 选取时，将具有好的稳定性质. 容易证明下面的定理.

定理 5.9 若 Q 由 (5.97) 选取,则公式 (5.66) 是 A 稳定的.

证明 令 $f(t,y)=\lambda y$, λ 是复数 $\mathrm{Re}\lambda<0$, 则由 (5.97) 确定的 q 为

$$q=-h\lambda,$$

并且有 $\phi_{n-i}=0$. 于是公式 (5.66) 给出

$$y_{n+1}=e^{\lambda h}y_n.$$

再由 $\mathrm{Re}\lambda<0$ 推出对任何固定的 h,当 $n\to\infty$ 时 $y_n\to 0$.

§4 一类特殊刚性方程的修正线性多步方法

在控制系统的设计中,经常会遇到形式为

$$P_N(D)y(t) = f(t, y) \tag{5.98}$$

的微分方程及某个给定的初始条件。$D = \frac{d}{dt}$ 是微分算子，$P_N(\cdot)$ 是具有离散根 $\lambda_i, i = 1,\cdots,N$ 的多项式,即有

$$P_N(\lambda) = \prod_{j=1}^{N} (\lambda - \lambda_j). \tag{5.99}$$

f 是 t 和 y 的任意光滑函数。 假定 $\left|\frac{\partial f}{\partial y}\right|$ 具有与 $\min_i |P'_N(\lambda_i)|$ 相当的量级。在 $|\lambda_i|$ 中可能有非常大的数。

现在应用形式为

$$\sum_{j=0}^{k} \alpha_j(h) y_{n+j} - \sum_{j=0}^{k} \beta_j(h) f_{n+j} = 0, \quad \alpha_k(h) = 1 \tag{5.100}$$

的广义线性多步方法，其中系数 $\alpha_j(h)$ 和 $\beta_j(h)(j = 1,\cdots,k)$ 依赖步长 h。y_{n+j} 为在 $t = t_n + jh$ 处的数值解。

$$f_{n+j} = f(t_n + jh, y_{n+j}).$$

令

$$\left.\begin{aligned}
\rho(\zeta, h) &= \sum_{j=0}^{k} \alpha_j(h)\zeta^j, \\
\sigma(\zeta, h) &= \sum_{j=0}^{k} \beta_j(h)\zeta^j, \\
L_h &= \rho(E, h) - \sigma(E, h)P_N(D), \\
L'_h &= E^{-k}L_h,
\end{aligned}\right\} \tag{5.101}$$

其中 E 是推移算子。

下面利用指数拟合的思想确定公式(5.100)，要求公式在点 $h\lambda_j(j = 1, 2,\cdots, m, m \geqslant N + 1)$ 处是 p_j 次指数拟合的,其中

$\lambda_j,\ j=1,\cdots,N$ 是多项式 P_N 的根，

$\lambda_{N+1}=0$,

$\lambda_j,\ j>N+1$ 是任意的，它们当方法与其它的多步公式联用时才需要确定. 由定义 5.4，当 $y(t)$ 有形式

$$y(t)=\sum_{j=1}^{m} r_j(t)e^{\lambda_j t}(m\geqslant N+1) \tag{5.102}$$

时，要求成立 $L_h y(t)=0$，其中 $r_j(t)$ 是次数不超过 p_j 的任意多项式. 这与当 $f(t,y(t))$ 有形式

$$f(t,y(t))=\sum_{j=1}^{m} q_j(t)e^{\lambda_j t} \tag{5.103}$$

时，方程 (5.98) 的解 $y(t)$ 精确地满足差分公式(5.100)是等价的，其中 $q_j(t)$ 是次数小于或等于 s_j 的任意多项式，

$$s_j=\begin{cases} p_j-1, & \text{如果 } P_N(\lambda_j)=0, \\ p_j & \text{其它.} \end{cases} \tag{5.104}$$

用

$$e_\lambda(t)=e^{\lambda t},\quad \phi_i(t)=t^i \tag{5.105}$$

定义函数 e_λ 和 ϕ_i，由多步算子 L_h 的线性性，系数 $\alpha_j(h)$ 和 $\beta_j(h),\ j=0,1,\cdots,k$ 之间的必要的关系式由下面的条件建立：

$$L_h(e_{\lambda_j}\phi_i)=0,\ j=1,2,\cdots,m,\ i=0,1,\cdots,p_j. \tag{5.106}$$

在 (5.106) 中，对 $j=1,2,\cdots,N$，令 $i=0$，得到

$$L_h(e_{\lambda_j})=0,\ j=1,\cdots,N. \tag{5.107}$$

按照 L_h 的定义式，得

$$[\rho(E,h)-\sigma(E,h)P_N(D)]e_{\lambda_j}=0,\ j=1,2,\cdots,N, \tag{5.108}$$

由于 $P_N(\lambda_j)=0,\ j=1,2,\cdots,N$. 从 (5.108) 推出

$$\rho(e^{\lambda_j h},h)=0,\ j=1,2,\cdots,N.$$

这就确定了 k 次多项式 ρ 的 N 个根. ρ 的其余 $k-N$ 个根可以任意选取. 但为了达到好的稳定性质，让它们等于零(类似于 Adams 方法). 因此，这一节的其余部分取 ρ 有形式

$$\rho(\zeta, h) = \zeta^k \prod_{j=1}^{N} (1 - e^{\lambda_j h} \zeta^{-1}). \tag{5.109}$$

由 (5.106),我们有

$$\sum_{j=1}^{m} (p_j + 1) - N = \sum_{j=1}^{m} (s_j + 1) \tag{5.110}$$

个条件来确定 k 次多项式 σ 的 $k+1$ 个系数. (5.110) 中的 N 表示已用掉 N 个条件来确定多项式 ρ. 这样,可以取多项式 σ 的次数 k 为

$$k = \sum_{j=1}^{m} (s_j + 1) - 1.$$

下面的定理给出我们需要的指数拟合的一个等价条件.

定理 5.10 令 L_h 是由 (5.101) 确定的算子, 则对所有 p 次多项式 φ_p,条件

$$L_h(\varphi_p e_\lambda) = 0 \tag{5.112}$$

等价于条件

$$L_h e_\mu = O((\mu - \lambda)^{p+1}) \quad (\mu \to \lambda). \tag{5.113}$$

证明 假定 $L_h(\varphi_p e_\lambda) = 0$. 由于 L_h 是线性算子,推出对所有 $i = 0, 1, \cdots, p$ 有 $L_h(\phi_i e_\lambda) = 0$. 由于有

$$\begin{aligned} L_h(\phi_i e_\lambda) &= L_h \left\{ \left[\frac{\partial^i}{\partial \mu^i} e_\mu \right]_{\mu=\lambda} \right\} \\ &= \left\{ L_h \left(\frac{\partial^i}{\partial \mu^i} e_\mu \right) \right\}_{\mu=\lambda}, \end{aligned} \tag{5.114}$$

应用 L_h 的定义,并且算子 $E, D, \dfrac{\partial}{\partial \mu}$ 作用到函数 e_μ 上是可交换的,我们得到

$$L_h(\phi_i e_\lambda) = \left\{ \frac{\partial^i}{\partial \mu^i} (L_h e_\mu) \right\} \Big|_{\mu=\lambda}. \tag{5.115}$$

于是由假定 $L_h(\phi_i e_\lambda) = 0 (i = 1, \cdots, p)$ 推得

$$\left\{ \frac{\partial^i}{\partial \mu^i} (L_h e_\mu) \right\}_{\mu=\lambda} = 0, \ i = 0, 1, \cdots, p, \tag{5.116}$$

将 $L_h e_\mu$ 绕点 λ 进行 Taylor 展开,由 (5.115) 得到

$$L_h e_\mu = O((\mu - \lambda)^{p+1}),$$

定理的逆的部分的证明是类似的。

在分析和实际计算中，有时将(5.101)中定义的算子 L'_h 改写成

$$L'_h = \prod_{j=1}^{N} (1 - e^{\lambda_j h} E^{-1}) - \left[\sum_{i=0}^{k} \delta_i(h) \nabla^i\right] P_N(D) \tag{5.117}$$

的形式可能比较方便。其中 $\nabla = I - E^{-1}$ 是向后差分算子。特别，令

$$P_N(D) = D, N = 1 \text{ 和 } \lambda_1 = 0, \tag{5.118}$$

得到 Adams-Moulton 公式。

为了确定算子 L'_h，只需确定(5.117)中的系数 $\delta_i(h)$，$i=0, 1, \cdots, k$。按照条件(5.106)和定理 5.10，有

$$L'_h e_\lambda = O[(\lambda - \lambda_\nu)^{p_\nu+1}], \nu = 1, 2, \cdots, m, \tag{5.119}$$

其中 $m \geqslant N+1$。应用公式(5.117)，给出

$$\left\{\left[\prod_{j=1}^{N} (1 - e^{\lambda_j h} E^{-1}) - \left(\sum_{i=0}^{k} \delta_i(h) \nabla^i\right) P_N(D)\right] e_\lambda\right\} = O[(\lambda - \lambda_\nu)^{p_\nu+1}]. \tag{5.120}$$

考虑到 $P_N(\lambda) = \prod_{j=1}^{N} (\lambda - \lambda_j)$，对这关系式作简单的变换得

$$\begin{aligned}\prod_{j=1}^{N} \frac{1 - e^{-(\lambda-\lambda_j)h}}{\lambda - \lambda_j} - \sum_{i=0}^{k} \delta_i(h)(1 - e^{-\lambda h})^i &= \begin{cases} O[(\lambda - \lambda_\nu)^{p_\nu}], & \text{如果 } P_N(\lambda_\nu) = 0 \\ O[(\lambda - \lambda_\nu)^{p_\nu+1}], & \text{其它} \end{cases} \\ &= O[(\lambda - \lambda_\nu)^{s_\nu+1}]\end{aligned} \tag{5.121}$$

令 $z = 1 - e^{-\lambda h}$ 和 $z_j = 1 - e^{-\lambda_j h}$，即 $\lambda = -\log(1-z)/h$ 和 $\lambda_j = -\log(1 - z_j)/h$，于是(5.121)有形式

$$\prod_{j=1}^{N} \frac{h}{1 - z_j} \frac{z - z_j}{\log(1 - z_j) - \log(1 - z)} - \sum_{i=0}^{k} \delta_i(h) z^i = O[(z - z_\nu)^{s_\nu+1}], \nu = 1, 2, \cdots, m, \tag{5.122}$$

其中

$$s_\nu=\begin{cases}p_\nu-1, & \text{如果 } P_N(\lambda_\nu)=0,\\ p_\nu, & \text{其它情形.}\end{cases}$$

于是求系数 $\delta_i(h)$ 的问题变成求 k 次多项式

$$\sum_{i=0}^{k}\delta_i(h)z^i$$

对函数

$$\phi(z)=\prod_{j=1}^{N}\frac{h}{1-z_j}\,\frac{z-z_j}{\log(1-z_j)-\log(1-z)}$$

的复密切插值问题,要求在 z_ν 处的插值阶为 s_ν.

上面讨论的是多项式 $P_N(\lambda)$ 是单根的情形.下面假定 $P_N(\lambda)$ 具有重根. 即假定 $P_N(D)$ 可表成

$$P_N(D)=\prod_{j=1}^{N}(D-\lambda_j)^{r_j}.$$

要求在 $h\lambda_j$, $j=1,\cdots,N$ 处作 $p_j\geqslant r_j-1$ 次的指数拟合. 这时多步算子 L_h' 为

$$L_h'=\prod_{j=1}^{N}(1-e^{\lambda_j h}E^{-1})^{r_j}-\left[\sum_{i=0}^{k}\delta_i(h)\nabla^i\right]P_N(D),$$

其中系数 $\delta_i(h)$ 可以这样确定: 求多项式 $\sum_{i=0}^{k}\delta_i(h)z^i$ 在 $z_\nu=1-e^{-\lambda_\nu h}$ 处对函数

$$\phi(z)=\prod_{j=1}^{N}\left[\frac{h}{1-z_j}\,\frac{z-z_j}{\log(1-z_j)-\log(1-z)}\right]^{r_j}$$

进行插值,对于 $\nu=1,\cdots,m$ 的插值阶为

$$s_\nu=\begin{cases}p_\nu-r_\nu & \text{如果 } P_N(\lambda_\nu)=0,\\ p_\nu & \text{其它.}\end{cases}$$

若 $P_N(\lambda)$ 的一些根相互很接近,可用其中的一个根来近似这些根,并在这个根上作指数拟合.

上面推导的多步公式的系数可以编制专门的程序来自动地计

算．当方程 (5.98) 的左边部分改变时，或者当步长 h 或某些参数 p_ν 变化时，系数都必须重新计算．这不会增加许多麻烦，因为对许多实际中遇到的问题 (5.98) 的左边部分和参数 h, p_ν 均是相对稳定的，计算不会太频繁．另一方面计算是自动进行的，并当 N 和 p_ν 不太大时，计算能很快完成．

上面推导的多步公式是为了想能拟合形式为

$$P_N(D)y(t) = \sum_{\nu=1}^{m} g_\nu(t)e^{\lambda_\nu t} \tag{5.123}$$

的微分方程的解，以便使我们的公式优于通常的方法．(5.123) 中的函数 $g_\nu(t)$ 与 e_{λ_ν} 相比，变化是比较缓慢的．下面来估计用构造的多步公式求微分方程

$$P_N(D)y(t) = g_\nu(t)e^{\lambda_\nu t}$$

的解 $y(t)$ 时的局部截断误差．

定义 5.6 算子 L'_h 应用到函数 $y(t)$ 时，称

$$[L'_h y](t_n)$$

为公式在 t_n 处的局部截断误差．

由 (5.101)，我们有

$$L'_h = \rho_1(E, h) - \sigma_1(E, h)P_N(D),$$

其中

$$\rho_1(\zeta, h) = \zeta^{-k}\rho(\zeta, h),$$
$$\sigma(\zeta, h) = \zeta^{-k}\sigma(\zeta, h).$$

令 $\mathscr{L}$ 是由

$$[\mathscr{L}y](\lambda) = \int_0^\infty e^{-\lambda t}y(t)dt$$

所定义的算子(单边 Laplace 变换)，其中存在一个实常数 x_1，使积分在半平面 $\mathrm{Re}\lambda > x_1$ 中是收敛的．由 Laplace 变换的逆公式给出

$$y(t_n) = \frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t_n}[\mathscr{L}y](\lambda)d\lambda, \quad x_2 > x_1.$$

现在可将在时刻 t_n 处的局部截断误差写成

$$[L'_h y](t_n) = \frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty}[L'_h e_\lambda](t_n)[\mathscr{L} y](\lambda)d\lambda,$$

由于

$$[L'_h e_\lambda](t_n) = e^{\lambda t_n}[\rho_1(e^{h\lambda}, h) - \sigma_1(e^{h\lambda}, h)P_N(\lambda)],$$

可以写出

$$[L'_h y](t_n) = \frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t_n}[\mathscr{L} F_1](\lambda)[\mathscr{L} F_2](\lambda)d\lambda, \tag{5.124}$$

其中

$$[\mathscr{L} F_1](\lambda) = \frac{\dfrac{\rho_1(e^{h\lambda}, h)}{P_N(\lambda)} - \sigma_1(e^{h\lambda}, h)}{(\lambda - \lambda_\nu)^{s_\nu+1}}, \tag{5.125}$$

$$[\mathscr{L} F_2](\lambda) = (\lambda - \lambda_\nu)^{s_\nu+1}P_N(\lambda)[\mathscr{L} y](\lambda). \tag{5.126}$$

由于构造的多步公式是在 $h\lambda_\nu$ 处指数拟合的，有

$$\frac{\rho_1(e^{h\lambda}, h)}{P_N(\lambda)} - \sigma_1(e^{h\lambda}, h) = O[(\lambda - \lambda_\nu)^{s_\nu+1}],$$

因此当 $\lambda \to \lambda_\nu$ 时，$[\mathscr{L} F_1](\lambda)$ 是有界的。

我们需要两个引理

引理 5.1 设上述的假定成立，则有

$$F_2(t) = g_\nu^{(s_\nu+1)}(t)e^{\lambda_\nu t}$$

证明 由 Laplace 变换的逆公式

$$\begin{aligned}
F_2(t) &= \frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t}[\mathscr{L} F_2](\lambda)d\lambda \\
&= \frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t}(\lambda - \lambda_\nu)^{s_\nu+1}P_N(\lambda)[\mathscr{L} y](\lambda)d\lambda \\
&= (D - \lambda_\nu)^{s_\nu+1}P_N(D)\frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t}[\mathscr{L} y](\lambda)d\lambda \\
&= (D - \lambda_\nu)^{s_\nu+1}P_N(D)y(t) = g_\nu^{(s_\nu+1)}(t)e^{\lambda_\nu t}.
\end{aligned}$$

引理 5.2 如果 $[\mathscr{L} F_1](\lambda)$ 由 (5.125) 定义，则对于 $u > kh$，有

$$F_1(u) \equiv 0.$$

这个引理的证明较繁，可参考 Bjurel [29]，这里省略掉。

我们还将用到卷积

$$(F_1 * F_2)(t) = \int_0^t F_1(u) F_2(t-u) du$$

的一个结果：

$$\mathscr{L} F_1 \cdot \mathscr{L} F_2 = \mathscr{L}(F_1 * F_2).$$

由这结果，应用 Laplace 变换的逆公式，给出

$$\frac{1}{2\pi i}\int_{x_2-i\infty}^{x_2+i\infty} e^{\lambda t_n}[\mathscr{L} F_1](\lambda)[\mathscr{L} F_2](\lambda) d\lambda = [F_1 * F_2](t_n),$$

所以，如果 $F_2 = g_\nu^{(s_\nu+1)} e_{\lambda_\nu}$，公式 (5.124) 变成

$$[L_h' y](t_n) = \int_0^{t_n} F_1(u) g_\nu^{(s_\nu+1)}(t_n - u) e^{\lambda_\nu(t_n-u)} du,$$

应用引理 5.2，得

$$[L_h' y](t_n) = e^{\lambda_\nu t_n}\int_0^{kh} e^{-\lambda_\nu u} F_1(u) g_\nu^{(s_\nu+1)}(t_n - u) du$$

$$= e^{\lambda_\nu t_n}\left[g_\nu^{(s_\nu+1)}(t_n)\int_0^\infty e^{-\lambda_\nu u} F_1(u) du + R\right], \quad (5.127)$$

其中

$$R = \int_0^{kh} e^{-\lambda_\nu u} F_1(u)[g_\nu^{(s_\nu+1)}(t_n - u) - g_\nu^{(s_\nu+1)}(t_n)] du. \quad (5.128)$$

因此

$$[L_h' y](t_n) = e^{\lambda_\nu t_n}[g_\nu^{(s_\nu+1)}(t_n)\lim_{\lambda\to\lambda_\nu}([\mathscr{L} F_1](\lambda)) + R], \quad (5.129)$$

可以这样来估计 R。首先应用 Cauchy-Schwartz 不等式，然后应用单边 Laplace 变换的 Parseval 关系式，得

$$|R| < \left[\int_0^\infty e^{-2\mathrm{Re}\lambda_\nu u}|F_1(u)|^2 du \int_0^{kh} |g_\nu^{(s_\nu+1)}(t_n - u) - g_\nu^{(s_\nu+1)}(t_n)|^2 du\right]^{\frac{1}{2}} = \left[\frac{1}{2\pi}\int_{\mathrm{Re}\lambda_\nu - i\infty}^{\mathrm{Re}\lambda_\nu + i\infty} |[\mathscr{L} F_1](\lambda)|^2 d\lambda \cdot \int_0^{kh} |g_\nu^{(s_\nu+1)}(t_n - u) - g_\nu^{(s_\nu+1)}(t_n)|^2 du\right]^{\frac{1}{2}}.$$

因此

$$|R| < \sqrt{\frac{kh}{2\pi}} \max_{0 \leqslant u \leqslant kh} |g_\nu^{(s_\nu+1)}(t_n - u)$$
$$- g_\nu^{(s_\nu+1)}(t_n)| \left[\int_{\mathrm{Re}\lambda_\nu - i\infty}^{\mathrm{Re}\lambda_\nu + i\infty} |[\mathscr{L} F_1](\lambda)|^2 dt\right]^{\frac{1}{2}}. \tag{5.130}$$

为了得到由公式(5.128)和(5.129)给出的局部截断误差的估计,我们必须估计 $[\mathscr{L} F_1](\lambda)$. 这个函数由(5.125)给出. 若 σ_1 的系数为一般的情形,它是未知的. 我们看到,按公式(5.121)和(5.122),有

$$(\lambda - \lambda_\nu)^{s_\nu+1}[\mathscr{L} F_1](\lambda) = \prod_{j=1}^{N} \frac{1 - e^{-(\lambda-\lambda_j)h}}{\lambda - \lambda_j}$$
$$- \sum_{i=0}^{k} \delta_i(h)(1 - e^{-\lambda h})^i = R_k(\phi, z), \tag{5.131}$$

其中 $z = 1 - e^{-\lambda h}$, $R_k(\phi, z)$ 是由(5.122)给出的插值的余项. 按照 Hermite 定理的推广,有: 如果 ϕ 在闭单连通区域 D 中是解析的, C 是位于 D 中的简单闭弧,并且在其内部含有不同的点 z_j, $j = 1, 2, \cdots, m$. 于是对于每个 $z \in D$

$$R_k(\phi, z) = \frac{1}{2\pi i}\int_C \frac{w(z)\phi(t)}{w(t)(t - z)} dt, \tag{5.132}$$

其中

$$w(z) = \prod_{j=1}^{m} (z - z_j)^{s_j+1}, \quad z_j = 1 - e^{-\lambda_j h}. \tag{5.133}$$

通过选取适当的积分路径,可以估计(5.132)中的积分. Bjurel 在[30]中导出局部截断误差的一个粗糙的估计: 对于实的 λ_i, $i = 1, 2, \cdots, N$,

$$|[L_h' y](t_n)| < e^{\mathrm{Re}\lambda_\nu t_n} e^{-h(s-1)\mathrm{Re}\lambda_\nu}[c_1 h^{s_\nu} |g_\nu^{(s_\nu+1)}(t_n)| + |R|] \tag{5.134}$$

(当 λ_ν 是实值时,可用 λ_ν 代替 $\mathrm{Re}\lambda_\nu$),其中

$$c_1 = 2^{N+1} \frac{1}{\min\limits_{1 \leqslant i \leqslant N} |\lambda_i|} \prod_{j=1}^{N} \frac{1}{|\lambda_j|},$$

$$s = \Sigma s_j < k,$$

$$|R| < c_2 h^{s_\nu+1} \|g_\nu^{(s_\nu+2)}\|_\infty,$$

$$c_2 = 2^{k+1-s_\nu} \sqrt{k}\ c_1.$$

数 s 中的求和是对所有有 $\mathrm{Re}\lambda_j > \mathrm{Re}\lambda_\nu$ 的 j 进行的. $|R|$ 的估计式中的最大模 $\|\cdot\|\infty$ 是在区间 $t_{n-kh} \leqslant t \leqslant t_n$ 上取的.

当函数 g_ν 满足某些正则化条件时，我们看到，当 $h \to 0$ 时，局部截断误差逼近于公式 (5.129) 中的第一项

$$e^{\lambda_\nu t_n} g_\nu^{(s_\nu+1)}(t_n)[\mathscr{L} F_1](\lambda_\nu).$$

当多步公式中的系数已知时，可以由公式 (5.125) 确定 $[\mathscr{L} F_1](\lambda)$，并计算局部截断误差的渐近表达式. 如果只要粗糙的估计，则 (5.134) 右边的第一项是可用的. 在 λ_j 是复的情形，除常数 c_1 和 c_2 可能稍大外，公式 (5.134) 仍成立.

在一般的情形，当 $h \to 0$ 时，公式 (5.125) 和 (5.129) 给出局部截断误差的渐近值

$$[L'_h y](t_n) \sim e^{\lambda_\nu t_n} g_\nu^{(s_\nu+1)}(t_n) \lim_{\lambda \to \lambda_\nu} \frac{\dfrac{\rho_1(e^{h\lambda}, h)}{P_N(\lambda)} - \sigma_1(e^{h\lambda}, h)}{(\lambda - \lambda_\nu)^{s_\nu+1}}. \tag{5.135}$$

下面是一个估计局部截断误差的简单例子

例 5.6 考虑微分方程

$$y' - qy = g(t), \quad \mathrm{Re}q < 0,$$

并且构造多步方法使在 hq 处是零次指数拟合的，而在原点是一次拟合的. 这时有

$$\rho_1(E, h) = 1 - e^{hq}E^{-1},$$

$$\sigma_1(E, h) = \frac{h}{(hq)^2}[(e^{hq} - 1 - hq) + ((hq - 1)e^{hq} + 1)E^{-1}].$$

按公式 (5.135)，有局部截断误差

$$[L'_h y](t_n) \sim g^{(2)}(t_n)\left[\left(1 - \frac{hq}{2}\right)e^{hq} - 1 - \frac{hq}{2}\right] \Big/ q^3, \tag{5.136}$$

这公式的右边可等价地记成为

$$h^3 g^{(2)}(t_n)[(1-hq/2)e^{hq}-1-hq/2]/(hq)^3.$$

于是当 $h\to 0$ 而 $hq\not\to 0$ 时，渐近地有

$$[E_h y](t_n)\sim\frac{h}{2q^2}g^{(2)}(t_n).$$

这时，由公式 (5.134) 的第一项给出的估计是小于 $\frac{8h}{q^2}g^{(2)}(t_n)$。

与§1中的公式 $F^{(1)}$ 和 (5.15) 相比，若记

$$g(t,y)=f(t,y)-qy,$$

$$\mu=\frac{1}{hq}\left(1-\frac{hq}{e^{hq}-1}\right).$$

上面构造的公式即是§1中的公式。(5.136) 给出这些公式当 $q\to-\infty$ 时的局部截断误差的渐近估计。

本 章 附 注

§1 中的材料主要取自 Liniger 和 Willoughby [80] 和 Liniger [78]。

§2 是根据 Mäkelä, Nevanlinna, Sipilä [81] 编写的。

§3 是根据 Sarkany 和 Liniger [102] 编写的。

§4 的材料主要取自 Bjurel [29],[30],[31]。

第六章 Richardson 外插方法

Dahlquist 利用 A 稳定的线性多步方法的经典性结果对方法的误差阶作了限制（见 § 2.2）．为了利用低阶 A 稳定方法计算出具有高阶计算误差的数值结果，Dahlquist[48] 建议采用梯形法的整体外插方法．计算实践表明，当计算开始时采用某种平滑过程，这种方法是求解刚性方程的一种有效的方法．它的理论基础为 Richardson 外插方法及解对步长的偶次幂的渐近展开式．另外，利用 Richardson 外插还可以构造一些别的算法．

§ 1 截断误差的渐近展开式

考虑常微分方程初值问题

$$\frac{dy}{dt} = f(t, y), \quad y(0) = z_0, \tag{6.1}$$

不失一般性，只考虑在区间 $[0, 1]$ 上解单个方程的情形，故有 $z_0 \in R$，$f(t, y) \in R$．假定 f 是 Lipschitz 连续的，则 (6.1) 的解 $y(t) \in C[0, 1]$，其中 $C[0, 1]$ 是以 $[0, 1]$ 为定义域，值域在 R 中的所有连续函数所组成的集合．定义赋范线性空间

$$E = \{y(t) \mid y(t) \in C^{(1)}[0, 1]\},$$

$$E^0 = \left\{\begin{pmatrix} d_0 \\ d \end{pmatrix} \,\middle|\, d_0 \in R,\ d(t) \in C[0, 1]\right\},$$

E 中元 y 的范数为 $\|y\|_E = \max\limits_{t\in[0,1]} |y(t)|$，$E^0$ 中元 $\begin{pmatrix} d_0 \\ d \end{pmatrix}$ 的范数为 $\left\|\begin{pmatrix} d_0 \\ d \end{pmatrix}\right\|_{E^0} = |d_0| + \max\limits_{t\in[0,1]} |d(t)|$．$C^{(1)}[0, 1]$ 是 $C[0, 1]$ 中所有连续可微的元素的集合．E 和 E^0 对于这样定义的范数都是 Banach

空间. 把 z_0 看成是定值, 定义由 E 到 E^0 中的映象 F: 对任何 $y\in E$, 令

$$Fy=\begin{pmatrix} y(0)-z_0 \\ y'(t)-f(t,y(t)) \end{pmatrix}\in E^0. \tag{6.2}$$

于是问题 (6.1) 可以改述成: 求 $z\in E$, 使有

$$Fz=0. \tag{6.3}$$

问题 (6.3) 是用 $\{E,\ E^0,\ F\}$ 来描述的. 我们将这个问题称作原始问题. 记作 $\mathscr{B}=\{E,\ E^0,\ F\}$.

问题 (6.3) 是一个无限维问题. 为了数值求解它, 通常采用离散化方法, 将原始问题换成一个有限维问题的序列. 一般这些有限维问题均可以构造性地求解. 随着有限维问题的维数增加, 序列向前推进, 就可以得到原始问题的愈来愈精确的近似解.

从数学上, 应用到给定原始问题 $\mathscr{B}=\{E,\ E^0,\ F\}$ 的离散化方法是指一个无限序列 $\mathscr{M}=\{E_n,\ E_n^0,\ \Delta_n,\Delta_n^0,\ \varphi_n\}_{n\in N'}$, 其中 E_n 和 E_n^0 均是有限维 Banach 空间, $\Delta_n: E\to E_n$ 和 $\Delta_n^0: E^0\to E_n^0$ 均是线性映射, 并且具有性质:

对于固定的 $y\in E$ 有 $\lim\limits_{n\to\infty}\|\Delta_n y\|_{E_n}=\|y\|_E$,

对于固定的 $d\in E^0$ 有 $\lim\limits_{n\to\infty}\|\Delta_n^0 d\|_{E_n^0}=\|d\|_{E^0}$.

$\varphi_n: (E\to E^0)\to(E_n\to E_n^0)$, 并且 F 包含在所有 φ_n 的定义域中. N' 是自然数集 N 的无限子集.

对于 $n\in N'$, 令 $F_n=\varphi_n(F)$, 我们得到问题 (6.3) 的离散化问题 $\mathscr{B}_n$: 求 $\zeta_n\in E_n$, 使有

$$F_n\zeta_n=0. \tag{6.4}$$

ζ_n 称作有限维问题 $\mathscr{B}_n$ 的解. 下面假定 E_n 和 E_n^0 的维数相同.

例 6.1 考虑用 Euler 方法求解问题 (6.1) 的离散化问题, 取 $N'=N$. 对于 $n\in N$, 令

$$G_n=\{\nu/n,\ \nu=0,\ 1,\cdots,\ n\},$$

$$E_n=(G_n\to R),\ \|\eta\|_{E_n}=\max_{\nu=0,1,\cdots,n}\left|\eta\left(\frac{\nu}{n}\right)\right|,$$

$$E_n^0 = (G_n \to R), \quad \|\delta\|_{E_n^0} = |\delta(0)| + \max_{\nu=1,\cdots,n} \left|\delta\left(\frac{\nu}{n}\right)\right|.$$

G_n 就是数值积分格点的集合. E_n 即是在格点集合上定义的函数所构成的空间. E_n^0 也是在格点集合上定义的函数所形成的空间，但将格点 0 上的值单独考虑. 按下面的方式定义映象 Δ_n, Δ_n^0, $\varphi_n(F) = F_n$. 对于 $y \in E$, 有

$$(\Delta_n y)\left(\frac{\nu}{n}\right) = y\left(\frac{\nu}{n}\right).$$

对于 $\begin{pmatrix} d_0 \\ d \end{pmatrix} \in E^0$, 有

$$\left(\Delta_n^0 \begin{pmatrix} d_0 \\ d \end{pmatrix}\right)\left(\frac{\nu}{n}\right) = \begin{cases} d_0, & \nu = 0, \\ d\left(\frac{\nu-1}{n}\right), & \nu = 1, 2, \cdots, n. \end{cases}$$

对于 $\eta \in E_n$, 有

$$[\varphi_n(F)\eta]\left(\frac{\nu}{n}\right)$$

$$= \begin{cases} \eta(0) - z_0, & \nu = 0, \\ \dfrac{\eta\left(\frac{\nu}{n}\right) - \eta\left(\frac{\nu-1}{n}\right)}{\frac{1}{n}} - f\left(\frac{\nu-1}{n}, \eta\left(\frac{\nu-1}{n}\right)\right), & \nu = 1, 2, \cdots, n. \end{cases}$$

这就是将 Fy 离散化后的形式. 问题 (6.4) 的唯一解 ζ_n 的存在性是显然的，因为 ζ_n 可由递推公式唯一确定.

为了使得上面的构造有意义，我们引进一些概念.

定义 6.1 应用到原始问题 $\mathscr{B}$ 的离散化方法 $\mathscr{M}$ 称作在 $y \in E$ 处与 $\mathscr{B}$ 是相容的: 如果 y 是在映象 F 和 $\varphi_n(F)\Delta_n$, $n \in N'$ 的定义域中，并且有

$$\lim_{\substack{n\to\infty \\ n \in N'}} \|\varphi_n(F)\Delta_n y - \Delta_n^0 F y\|_{E_n^0} = 0.$$

(在以后对所有 $n \to \infty$ 时 n 的约束 $n \in N'$ 均省略掉), $\mathscr{M}$ 称作与 $\mathscr{B}$ 相容的，如果它在每个 $y \in E$ 处均与 $\mathscr{B}$ 相容. $\mathscr{M}$ 称作在 y 处

与 $\mathscr{B}$ 是 p 阶相容的，如果当 $n\to\infty$ 时有

$$\|\varphi_n(F)\Delta_n y-\Delta_n^0 F y\|_{E_n^0}=O(n^{-p}).$$

定义 6.2 令 $\mathscr{M}$ 与 $\mathscr{B}$ 在问题 (6.3) 的真解 z 处是相容的，则 $l_n=\varphi_n(F)\Delta_n z$, $n\in N'$ 称作 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部离散化误差.

粗略地说，l_n 是将 z 离散化后代入算子 F_n 所得的值.

定义 6.3 设 $\mathscr{B}$ 具有真解 z，离散化 $\mathscr{M}$ 具有唯一解序列 $\{\zeta_n\}_{n\in N'}$(ζ_n 由 (6.4) 确定)，则由

$$\varepsilon_n=\zeta_n-\Delta_n z$$

所确定的序列 $\{\varepsilon_n\}_{n\in N'}$ 称作 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散化误差. 如果有 $\lim\limits_{n\to\infty}\|\varepsilon_n\|_{E_n}=0$，则 $\mathscr{M}$ 称作对 $\mathscr{B}$ 是收敛的. 如果有 $\|\varepsilon_n\|_{E_n}=O(n^{-p})(n\to\infty)$，则称 $\mathscr{M}$ 对 $\mathscr{B}$ 是 p 阶收敛的.

ε_n 可以看成是用 $F_n\zeta_n=0$ 代替 $F_n\zeta_n=l_n$ 求解时二者之间的差. 为了考虑 l_n 对 ζ_n 的影响，引进下面离散化误差对于离散化解的稳定性概念.

定义 6.4 离散化 $\mathscr{M}$ 称作对序列 $\eta=\{\eta_n\}_{n\in N'}$, $\eta_n\in E_n$ 是稳定的，如果存在常数 s 和 $r>0$，使对所有 $n\in N'$ 一致地有：对所有满足 $\|F_n\eta_n^{(i)}-F_n\eta_n\|_{E_n^0}<r$ 的 $\eta_n^{(i)}(i=1,2)$，成立估计式

$$\|\eta_n^{(1)}-\eta_n^{(2)}\|_{E_n}\leqslant s\|F_n\eta_n^{(1)}-F_n\eta_n^{(2)}\|_{E_n^0}.$$

常数 s 和 r 分别称作稳定性的界和限. 如果 $\mathscr{M}$ 对序列 $\{\Delta_n z\}$ 是稳定的，则 $\mathscr{M}$ 称作对 $\mathscr{B}$ 是稳定的.

这个定义是说，对于在 $\{\eta_n\}$ 附近的 $\{\eta_n^{(1)}\}$ 和 $\{\eta_n^{(2)}\}$，可以用 $\{F_n\eta_n^{(i)}\}_{i=1,2}$ 之间的差来估计 $\{\eta_n^{(i)}\}_{i=1,2}$ 之间的差.

定义 6.5 如果映象 $\Lambda_n:E\to E^0$, $n\in N'$ 对于 F 的定义域中的所有 y 有

$$\varphi_n(F)\Delta_n y=\Delta_n^0(F+\Lambda_n)y, \tag{6.5}$$

则称它为 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差映象.

粗略地说，若不管空间不同，Λ_n 是两个算子 F_n 与 F 之差.

定义 6.6 如果存在与 n 无关的 E 的非空子集 D_J 和映象 λ_j: $D_J\to E^0$, $j=1,\cdots,J$，使对所有 $y\in D_J$ 有

$$\left\|\Delta_n^0\left[\Lambda_n y-\sum_{j=1}^{J}\frac{1}{n^j}\lambda_j y\right]\right\|=O(n^{-(J+1)}),\ n\to\infty,\ n\in N',$$
(6.6)

则称 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差映象 Λ_n 具有到阶 J 的渐近展开

$$\sum_{j=1}^{J}\frac{1}{n^j}\lambda_j.$$

λ_j 与 n 的无关性使在所有通常的应用中除掉阶为 $O(n^{-(J+1)})$ 的项外，Λ_n 是唯一确定的。于是如果 $y\in D_J$,则有

$$F_n\Delta_n y=\Delta_n^0\left[Fy+\sum_{j=1}^{J}\frac{1}{n^j}\lambda_j y\right]+O(n^{-(J+1)}),\qquad(6.7)$$

特别,如果 $\mathscr{B}$ 的真解 $z\in D_J$,则有

$$l_n=F_n\Delta_n z=\sum_{j=1}^{J}\frac{1}{n^j}\Delta_n^0(\lambda_j z)+O(n^{-(J+1)}).\qquad(6.8)$$

(6.8) 式是 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差 $\{l_n\}$ 的到阶 J 的渐近展开。如果 $\mathscr{M}$ 对 $\mathscr{B}$ 是 p 阶相容的,则有

$$\lambda_j z=0,\ j=1,2,\cdots,p-1.$$

例 6.2　继续考虑例 6.1. 由于有

$$[(F_n\Delta_n-\Delta_n^0F)y]\left(\frac{\nu}{n}\right)$$

$$=\begin{cases}0, & \nu=0,\\ \dfrac{y\left(\frac{\nu}{n}\right)-y\left(\frac{\nu-1}{n}\right)}{\frac{1}{n}}-y'\left(\frac{\nu-1}{n}\right), & \nu=1,2,\cdots,n.\end{cases}\qquad(6.9)$$

于是对每个 $n\in N$,映象

$$\Lambda_n:y\to\begin{pmatrix}0\\ \dfrac{y\left(t+\frac{1}{n}\right)-y'(t)}{\frac{1}{n}}-y'(t)\end{pmatrix}\qquad(6.10)$$

是局部误差映象．对于在格点 $\frac{\nu}{n}$ 上为零的任何项均可加到 Λ_n 上，但得到的仍是局部误差映象．

对于每个 $J\in N$，由 (6.10) 定义的 $\{\Lambda_n\}_{n\in N}$ 具有到 J 的渐近展开，其中 $D_J = C^{(J+2)}[0, 1]$，映象 λ_j 为

$$\lambda_j: y \to \begin{pmatrix} 0 \\ \frac{1}{(j+1)!} y^{(j+1)}(t) \end{pmatrix}, \quad j = 1, \cdots, J.$$

由于当 $y\in D_J$ 时有展开式

$$\frac{y\left(\frac{\nu}{n}\right) - y\left(\frac{\nu-1}{n}\right)}{\frac{1}{n}} = y'\left(\frac{\nu-1}{n}\right)$$

$$+ \sum_{j=1}^{J} \frac{1}{n^j} \frac{y^{(j+1)}\left(\frac{\nu-1}{n}\right)}{(j+1)!} + O(n^{-(J+1)}).$$

如果 $z\in D_J$，局部离散化误差 $\{l_n\}$ 具有渐近展开

$$l_n\left(\frac{\nu}{n}\right)$$

$$= \begin{cases} 0, & \nu = 0 \\ \sum_{j=1}^{J} \frac{1}{n^j} \frac{z^{(j+1)}\left(\frac{\nu-1}{n}\right)}{(j+1)!} + O(n^{-(J+1)}), & \nu = 1, 2, \cdots, n. \end{cases}$$

对于 $J = 1$，这归结成

$$l_n\left(\frac{\nu}{n}\right) = \frac{1}{2n} z''\left(\frac{\nu-1}{n}\right) + O(n^{-2}), \quad \nu = 1, 2, \cdots, n.$$

例 6.3　对方程 $y' = f(y)$ 的梯形法，定义

$$[\varphi_n(F)\eta]\left(\frac{\nu}{n}\right)$$

$$
= \begin{cases} \eta(0) - z_0 - \sum_{\nu=1}^{M} q_\nu \left(\frac{1}{n}\right)^{2\nu} + O\left(\left(\frac{1}{n}\right)^{2M+1}\right), \\ \dfrac{\eta\left(\frac{\nu}{n}\right) - \eta\left(\frac{\nu-1}{n}\right)}{\frac{1}{n}} - \frac{1}{2}\left[f\left(\eta\left(\frac{\nu}{n}\right)\right)\right. \\ \qquad \left. + f\left(\eta\left(\frac{\nu-1}{n}\right)\right)\right], \quad \nu = 1, \cdots, n, \end{cases}
$$

其中 q_ν 是实数. 如果由

$$
\left[\Delta_n^0 \binom{d_0}{d}\right]\left(\frac{\nu}{n}\right) = \begin{cases} d_0, & \nu = 0, \\ d\left(\frac{2\nu-1}{2n}\right), & \nu = 1, 2, \cdots, n \end{cases}
$$

定义 Δ_n^0,而由

$$
\Lambda_n : y \to \begin{cases} -\sum_{\nu=1}^{M} q_\nu \left(\frac{1}{n}\right)^{2\nu} + O\left(\left(\frac{1}{n}\right)^{2M+1}\right), \\ \dfrac{y\left(t+\frac{1}{2n}\right) - y\left(t-\frac{1}{2n}\right)}{\frac{1}{n}} - y'(t) \\ \quad - \frac{1}{2}\left[f\left(y\left(t+\frac{1}{2n}\right)\right) + f\left(y\left(t-\frac{1}{2n}\right)\right)\right] + f(y(t)) \end{cases} \tag{6.11}
$$

定义局部误差映象. 由 (6.11) 式,除掉阶为 $2M+1$ 的量外, $\Lambda_n y$ 对 n 显然是偶的. 于是在 $\Lambda_n y$ 的渐近展开式中,对所有小于 $2M+1$ 的奇数 j, 一定有 $\lambda_j = 0$.

定义 6.7 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散化误差 $\{\varepsilon_n\}_{n\in N'}$ 称作具有到阶 J 的渐近展开,如果存在与 n 无关的元 $e_j \in E$, $j = 1, \cdots, J$, 使有

$$
\varepsilon_n = \Delta_n \left[\sum_{j=1}^{J} \frac{1}{n^j} e^j\right] + O(n^{-(J+1)}). \tag{6.12}
$$

显然, 如果 $\mathscr{M}$ 对 $\mathscr{B}$ 是 p 阶收敛的(定义 6.3), 则 $e_j = 0$, $j = 1, \cdots, p-1$,而 $e_p \neq 0$.

关于到阶 $J \geqslant p$ 的整体离散化误差的渐近展开式的存在性的知识具有实际的重要性，而局部离散误差的渐近展开主要是作为分析整体离散误差的工具.

下面需要某些关于局部误差映象 $\{\Lambda_n\}$ 的渐近展开的光滑性条件.

定义 6.8 局部误差映象 $\{\Lambda_n\}_{n\in N'}$ 的到阶 J 的渐近展开 (6.6) 称作是 (J, p) 光滑的, $J \geqslant p$, 如果等式

$$\sum_{j=1}^{J} \frac{1}{n^j}\left[\lambda_j y + \sum_{m=1}^{\left[\frac{J-j}{p}\right]} \frac{1}{m!}\lambda_j^{(m)}(y)\left(\sum_{k=p}^{J}\frac{1}{n^k}e_k\right)^m\right]$$

$$= \Lambda_n\left(y + \sum_{k=p}^{J}\frac{1}{n^k}e_k\right) + O(n^{-(J+1)}) \qquad (6.13)$$

左边的导数存在,并且对任意 $y \in D_J$, $e_k \in D_J$, $k=p, p+1, \cdots, J$ (6.13) 式均成立.

这里 $\lambda_j^{(m)}$ 是 λ_j 的 m 阶导算子(导数), 所用到的均假定其存在. $\sum_{k=p}^{J}\frac{1}{n^k}e_k$ 是 $\mathscr{M}$ 对 $\mathscr{B}$ 在 p 阶收敛时 ε_n 在 E 中的原象(误差在 $O(n^{-(J+1)})$ 的范围内). 定义 6.8 是说: 若 Λ_n 作用于 $y+\sum_{k=p}^{J}\frac{1}{n^k}e_k$ 时,将 Λ_n 换成其渐近展开 $\sum_{j=1}^{J}\frac{1}{n^j}\lambda_j$ 之后,又可将 λ_j 的每项展开,其误差在 $O(n^{-(J+1)})$ 的范围之内前后相等. 在左端方括号内的和中, m 最大到 $\left[\frac{J-j}{p}\right]$, 因为 ε_n 的原象中 $\frac{1}{n}$ 最小的阶为 p, 当 $m=\left[\frac{J-j}{p}\right]+1$ 时, 这项的阶是 $p\left(\left[\frac{J-j}{p}\right]+1\right)+j \geqslant J+1$, 而当 $m=\left[\frac{J-j}{p}\right]$ 时, $p\left[\frac{J-j}{p}\right]+j \leqslant J$, 所以 m 最大到 $\left[\frac{J-j}{p}\right]$ 就保证误差在 $O(n^{-(J+1)})$ 之内.

如果渐近展开 (6.6) 是 (J, p) 光滑的, 用集合 $D_{J,k} \subset E$, $k=p, p+1, \cdots, J$ 表示使 (6.13) 对 $y \in D_J$, $e_k \in D_{J,k}$, $k=p$, $p+$

$1, \cdots, J$ 均成立的任意区域.

定理 6.1 设 $\mathscr{B}$ 具有真解 z, $\mathscr{M}$ 为应用到 $\mathscr{B}$ 的离散化方法. 假定 $\mathscr{M}$ 和 $\mathscr{B}$ 满足下面各条件

(i) $\mathscr{M}$ 对 $\mathscr{B}$ 是稳定的.

(ii) $\mathscr{M}$ 对 $\mathscr{B}$ 是 p 阶相容的. $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差映象存在,并且有到阶 J 的渐近展开式, $z \in D_J$.

(iii) 在某个球 $B_R = \{y \in E \mid \|y - z\| < R\}$ 中, F 具有 $[J/p]$ 阶的 Lipschitz 连续的 Fréchet 导数, 并且 (ii) 中的渐近展开式是 (J, p) 光滑的.

(iv) $F'(z)^{-1}$ 存在.

对于 $j = 2p, 2p+1, \cdots, J$, 定义映象 $g_j: D_{J,p} \times D_{J,p+1} \times \cdots \times D_{J,j-p} \to E^0$, 使等式

$$\sum_{j=2p}^{J} \frac{1}{n^j} g_j(e_p, \cdots, e_{j-p})$$

$$= \sum_{m=2}^{[J/p]} \frac{1}{m!} \left[F^{(m)}(z) + \sum_{j=1}^{J-mp} \frac{1}{n^j} \lambda_j^{(m)}(z) \right]$$

$$\cdot \left(\sum_{k=p}^{J} \frac{1}{n^k} e_k \right)^m + O(n^{-(J+1)}) \qquad (6.14)$$

成立. 对于 $j = p, p+1, \cdots, 2p-1$ 定义 $g_j \equiv 0$. 如果由式

$$F'(z)e_j = -\left[\lambda_j z + \sum_{k=1}^{j-p} \lambda_k'(z) e_{j-k} + g_j(e_p, \cdots, e_{j-p}) \right] \qquad (6.15)$$

可以逐次定义 $e_j \in E$, $j = p, p+1, \cdots, J$, 并且有

$$e_j \in D_{J,j}, \quad j = p, \ p+1, \cdots, J. \qquad (6.16)$$

于是 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散误差 $\{\varepsilon_n\}_{n \in N'}$ 具有到阶 J 的唯一的渐近展开

$$\varepsilon_n = \Delta_n \left(\sum_{j=p}^{J} \frac{1}{n^j} e_j \right) + O(n^{-(J+1)}). \qquad (6.17)$$

证明 对于上面定义的 e_j, 考虑公式

$$F\left(z+\sum_{k=p}^{J}\frac{1}{n^k}e_k\right)+\Lambda_n\left(z+\sum_{k=p}^{J}\frac{1}{n^k}e_k\right)$$

$$=F'(z)\sum_{k=p}^{J}\frac{1}{n^k}e_k+\sum_{i=p}^{J}\frac{1}{n^i}\lambda_i z+\sum_{i=1}^{J}\frac{1}{n^i}\lambda_i'(z)$$

$$\cdot\left(\sum_{k=p}^{J-i}\frac{1}{n^k}e_k\right)+\sum_{m=2}^{[J/p]}\frac{1}{m!}\Bigg[F^{(m)}(z)$$

$$+\sum_{i=1}^{J-mp}\frac{1}{n^i}\lambda_i^{(m)}(z)\Bigg]\left(\sum_{k=p}^{J}\frac{1}{n^k}e_k\right)^m$$

$$+R_J(z,e_p,\cdots,e_J). \tag{6.18}$$

这里当 $r>s$ 时令 $\sum_{k=r}^{s}(\quad)=0$. 按照假定(iii)和(6.16),有

$$\|R_J(z,e_p,\cdots,e_J)\|=O(n^{-(J+1)}). \tag{6.19}$$

由于对于充分大的 n, $z+\sum_{k=p}^{J}\frac{1}{n^k}e_k\in B_R$,由(6.14),(6.15),(6.18)的右边的各项(除了阶为 $O(n^{-(J+1)})$ 的项外)的和为零. 由于(6.15)构造 e_i 即是为了达到这个目的. 这样,我们推出

$$F_n\Delta_n\left(z+\sum_{k=p}^{J}\frac{1}{n^k}e_k\right)=O(n^{-(J+1)}), \tag{6.20}$$

由 ε_n 的定义, $F_n(\Delta_n z+\varepsilon_n)=0$, 则从(i)推出

$$\varepsilon_n=\Delta_n\left(\sum_{j=p}^{J}\frac{1}{n^j}e_j\right)+O(n^{-(J+1)}). \tag{6.21}$$

唯一性由离散化方法的映象 Δ_n 的性质 $\lim_{n\to\infty}\|\Delta_n y\|_{E_n}=\|y\|_E$ 推出. 事实上,假定存在另外一个展开式

$$\varepsilon_n=\Delta_n\left(\sum_{j=p}^{J}\frac{1}{n^j}\hat{e}_j\right)+O(n^{-(J+1)}), \tag{6.22}$$

其系数 $\hat{e}_j\in E$ 与 n 无关. 将(6.21)和(6.22)相减,并乘上 n^p,得到

$$\Delta_n(\hat{e}_p-e_p)=O\left(\frac{1}{n}\right).$$

令 $n\to\infty$ 可推得 $\hat{e}_p=e_p$. 按同样的方式可递推地得到 $\hat{e}_j=e_j$, $j=p+1, p+2, \cdots, J$.

e_i 是由原始问题定义的，它由问题 (6.15) 来确定. 离散化 $\mathscr{M}$ 是通过 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差映象的渐近展开式的系数 $\lambda_i\in[E\to E^0]$ 在 (6.15) 中体现. 对于给定的 $\mathscr{M}$ 和 $\mathscr{B}$, 验证定理 6.1 的假定的主要工作是验证局部误差映象的渐近展开式的 (J, p) 光滑性和对适当选取的集合 $D_{J,j}$, 条件 (6.16) 是否成立. 上面的定理具有构造的性质，指出了如何去构造渐近展开式的系数.

例 6.4　继续研究例 6.1、例 6.2. 假定 (i) 和 (iv) 满足，具有 1 阶的相容性. 例 6.2 中在 $D_J=C^{(J+1)}[0, 1]$ 上给出了具有阶 J 的渐近展开式的局部误差映象. 出现在假定 (iii) 中的导数为

$$F'(y)e=\begin{pmatrix} e(0) \\ e'(t)-f'_y(t, y(t))e(t)\end{pmatrix},$$

$$F^{(m)}(y)e^m=\begin{pmatrix} 0 \\ -f_y^{(m)}(t, y(t))e(t)^m\end{pmatrix}.$$

对于 $j=1, 2, \cdots, J$

$$\lambda_j'(y)e=\begin{pmatrix} 0 \\ \dfrac{1}{(j+1)!}e^{(j+1)}(t)\end{pmatrix}=\lambda_j e,$$

$$\lambda_j^{(m)}(y)e^m=\begin{pmatrix} 0 \\ 0\end{pmatrix}, \quad m\geqslant 2.$$

因此，我们必须要求 f 在 $z(t)$, $t\in[0, 1]$ 的邻域中具有对 y 的 $J+1$ 阶连续偏导数. 由于 Λ_n 和 λ_j 的线性性，对 $y, e_1, \cdots, e_J\in D_J$, 有(只写出第二个分量)

$$\begin{aligned}\Lambda_n\left(y+\sum_{k=1}^J\frac{1}{n^k}e_k\right)&=\sum_{j=1}^J\frac{1}{n^j}\lambda_j\left(y+\sum_{k=1}^J\frac{1}{n^k}e_k\right)\\&\quad+O(n^{-(J+1)})=\sum_{j=1}^J\frac{1}{n^j}\left[\lambda_j y+\lambda_j\left(\sum_{k=1}^J\frac{1}{n^k}e_k\right)\right]\\&\quad+O(n^{-(J+1)}).\end{aligned}$$

对于 $e_k\in D_{J,k}=D_{J-k}$, $k=1,2,\cdots,J$,这关系式仍成立. 这样就验证了 $(J,1)$ 光滑性. 对于 (6.14),其中各项第一列都是零,由第二列,我们得到(仍以 g_i 表示原来的 g_i 的第二列)

$$\sum_{j=2}^{J}\frac{1}{n^j}g_j(e_1,\cdots,e_{j-1})=-\sum_{m=2}^{J}\frac{1}{m!}f_y^{(m)}(z(t))$$

$$\cdot\left(\sum_{k=1}^{J}\frac{1}{n^k}e_k\right)^m+O(n^{-(J+1)}).$$

所以有

$$g_1=0,$$

$$g_2(e_1)=-\frac{1}{2}f_y''(t,z(t))e_1^2,$$

$$g_3(e_1,e_2)=-f_y''(t,z(t))\,e_1e_2-\frac{1}{6}f'''(t,z(t))e_1^3,$$

$$g_4(e_1,e_2,e_3)=-\frac{1}{2}f''(t,z(t))(2e_1e_3+e_2^2)$$

$$-\frac{1}{2}f'''(t,z(t))e_1^2e_2-\frac{1}{24}f^{(IV)}(t,z(t))e_1^4,$$

$$\cdots\cdots$$

于是由 (6.15),得到确定 e_i, $i=1,2,\cdots,J$ 的递推式

$$e_i(0)=0,$$

$$e_i'(t)-f_y'(t,z(t))e_i(t)=b_i(t),\ t\in[0,1],$$

其中

$$b_j=-\frac{1}{(j+1)!}z^{(j+1)}(t)$$

$$-\sum_{k=1}^{j-1}\lambda_k e_{j-k}-g_j(e_1,\cdots,e_{j-1}).$$

因而

$$b_1=-\frac{1}{2}z'',$$

$$b_2=-\frac{1}{6}z'''-\frac{1}{2}e_1''+\frac{1}{2}f_y''(t,z)e_1^2,$$

$$b_3 = -\frac{1}{24}z^{(\mathrm{IV})} - \frac{1}{6}e_1''' - \frac{1}{2}e_2''$$

$$+ f_y''(t, z)e_1 e_2 + \frac{1}{6}f_y'''(t, z)e_1^3.$$

$$\cdots$$

由 $z \in D_J$，递推地可得到 $e_i \in D_{J-i}$. 因此推出这些 e_i 是对问题 (6.1) 的 Euler 方法的整体离散误差的渐近展开式的系数.

如果 $\mathscr{M}$ 对原始问题 $\mathscr{B}$ 的局部误差映象的渐近展开式中只含有 $\frac{1}{n}$ 的偶次幂，则在整体误差的渐近展开式中也具有这个性质. 有下面的定理.

定理 6.2 在定理 6.1 中，如果 $\mathscr{M}$ 对 $\mathscr{B}$ 的局部误差映象的渐近展开式中只含有 $\frac{1}{n}$ 的偶次幂，即

$$\lambda_i y = 0, 对\ y \in D_J,\ i\ 是奇数.$$

则 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散误差的渐近展开也是 $\frac{1}{n}$ 的偶次幂的，即若 i 是奇数，有

$$e_i = 0.$$

证明 只要在定理 6.1 的证明中将对应于奇数 i 的与 λ_i 有关的项看成零，即可得到定理的证明.

例 6.5 继续例 6.3 的讨论. 类似于 Euler 方法，对于 $p = 2$ 的梯形法也可以验证定理 6.1 中的假定成立. 再由例 6.3 中 Λ_n 的渐近展开式只含有 $\frac{1}{n}$ 的偶次幂. 应用定理 6.2，推出梯形法的整体离散误差仅含 $\frac{1}{n}$ 的偶次幂.

对于单步方法，Stetter 建立了下面的结果. 求解问题 (6.1) 的单步法可以写成形式

$$y_\nu = y_{\nu-1} + h\Phi(t_{\nu-1}, y_{\nu-1}, h),\ y_0 = z_0, \tag{6.23}$$

其中 $\Phi(t, y, h)$ 称作单步方法的增量函数，它由 $f(t, y)$ 和方法

所确定，仅是 t_ν, y_ν, h 的函数. 下面假定增量函数 $\Phi(t, y, h)$ 对 y 满足 Lipschitz 条件，对 y, t, h 具有所需要的一切可微性和连续性，并且有

$$\Phi(t, y(t), 0) = f(t, y(t)). \tag{6.24}$$

有下面的定理.

定理 6.3 如果增量函数 Φ 满足

$$\Phi(t + h, y + h\Phi(t, y, h), -h) = \Phi(t, y, h) \tag{6.25}$$

则单步方法 (6.23) 的解的整体离散化误差具有 $h = \frac{1}{n}$ 的偶次幂的渐近展开.

证明 对于 (6.1) 的近似解 y_ν, $\nu = 0, 1, \cdots$, 由 (6.25) 推出

$$\Phi(t_\nu, y_\nu, -h) = \Phi(t_{\nu-1}, y_{\nu-1}, h). \tag{6.26}$$

(6.23) 等价于

$$\frac{y_\nu - y_{\nu-1}}{h} = \frac{1}{2}[\Phi(t_\nu, y_\nu, -h) + \Phi(t_{\nu-1}, y_{\nu-1}, h)]. \tag{6.27}$$

令 $y(t)$ 是 (6.1) 的解，且 $\hat{t}_\nu = t_{\nu-1} + \frac{h}{2} = t_\nu - \frac{h}{2}$. (6.27) 的局部离散误差满足

$$\begin{aligned} l(\hat{t}_\nu, h) &= \frac{y(t_\nu) - y(t_{\nu-1})}{h} - \frac{1}{2}[\Phi(t_\nu, y(t_\nu), -h) \\ &\quad + \Phi(t_{\nu-1}, y(t_{\nu-1}), h)] \\ &= \frac{y\left(\hat{t}_\nu + \frac{h}{2}\right) - y\left(\hat{t}_\nu - \frac{h}{2}\right)}{h} \\ &\quad - \frac{1}{2}\left[\Phi\left(\hat{t}_\nu + \frac{h}{2}, y\left(\hat{t}_\nu + \frac{h}{2}\right), -h\right)\right. \\ &\quad \left. + \Phi\left(\hat{t}_\nu - \frac{h}{2}, y\left(\hat{t}_\nu - \frac{h}{2}\right), h\right)\right] = l(\hat{t}_\nu, -h). \end{aligned}$$

因此，与例 6.3 一样，局部离散误差按 h 的幂的渐近展开仅含 h 的

偶次幂. 由 Gear[20] 中的定理 4.2 和定理 4.3 易知定理 6.2 的条件成立，从而由定理 6.2 推得所需要的结论.

应用条件 (6.25) 通常是困难的，因为它对 Φ 是隐式的，而且通常 Φ 也是隐式地定义的. 在许多应用中，条件 (6.25) 改成下面的形式应用起来可能比较方便.

推论 6.1 如果单步算法 (6.23) 可以写成形式

$$\frac{y_\nu - y_{\nu-1}}{h} = \phi(t_\nu, t_{\nu-1}, y_\nu, y_{\nu-1}, h), \tag{6.28}$$

其中 ϕ 满足

$$\phi(s, t, \eta, \zeta, h) = \phi(t, s, \zeta, \eta, -h), \tag{6.29}$$

则定理 6.3 的结论成立.

条件 (6.25) 或条件 (6.29) 的直观推论是：如果 y_ν 是由 $y_{\nu-1}$ 通过增量 $h\Phi(t_{\nu-1}, y_{\nu-1}, h)$ 得到的，则 $y_{\nu-1}$ 可以由同一个增量函数 Φ 从 y_ν 通过增量 $-h\Phi(t_\nu, y_\nu, -h)$ 得到. 这就是方法所具有的某种对称性. 它在 (6.26), (6.27), (6.29) 中是很明显的.

例 6.6 隐式中点单步方法

$$y_\nu = y_{\nu-1} + hf\left(\frac{t_\nu + t_{\nu-1}}{2}, \frac{y_\nu + y_{\nu-1}}{2}\right)$$

所对应的 $\phi(t_\nu, t_{\nu-1}, y_\nu, y_{\nu-1}, h)$ 为

$$\phi(t_\nu, t_{\nu-1}, y_\nu, y_{\nu-1}, h) = f\left(\frac{t_\nu + t_{\nu-1}}{2}, \frac{y_\nu + y_{\nu-1}}{2}\right),$$

显然满足推论的条件.

梯形法

$$y_\nu = y_{\nu-1} + \frac{h}{2}[f(t_\nu, y_\nu) + f(t_{\nu-1}, y_{\nu-1})]$$

所对应的 ϕ 为

$$\phi(t_\nu, t_{\nu-1}, y_\nu, y_{\nu-1}, h) = \frac{1}{2}[f(t_\nu, y_\nu) + f(t_{\nu-1}, y_{\nu-1})],$$

也满足推论 1 的条件.

§ 2 Richardson 外插方法

在数值分析的各个领域中，例如数值微分，数值积分．方程求根等，经常会遇到这样的计算过程：要计算的量为 ζ_0，而我们只能计算其近似值 ζ_n，其中 $n\in N'\subset N$，N 是自然数集合，当 $n\to\infty$ 时，$\zeta_n\to\zeta_0$．这样的过程的收敛速度可能是比较慢的．希望能利用得到的 ζ_n 来加速这个极限过程． Richardson 外插方法就是这样一种方法． 它的基本思想是将收敛的离散化解 ζ_n 看成是 n 的函数，计算这个函数的少量的值 ζ_{n_ρ}，并利用这些值构造一个自变量为 n 的插值函数 $\chi(n)$． 于是原始问题的真解 ζ_0 的极限过程由取插值函数 $\chi(n)$ 在 $n=\infty$ 处的值来模拟．Richardson 称这种过程为"对极限的延伸处理"．正是他第一个应用这种思想来提高近似方法得到的结果的精度．

在数值求解常微分方程初值问题 (6.1) 时，选取一个严格单调收敛到零的正步长的序列 $\{h_n\}_{n\in N'}$．设用某种给定的数值方法以 h_n 为步长得到的解为 ζ_n；则当 $n\to\infty$ 时，在一定的意义下 ζ_n 将收敛到 (6.1) 的解 ζ_0．由 § 1，若 ζ_n 的整体离散化误差具有渐近展开，则在某种意义下，ζ_n 可表成

$$\zeta_n=\zeta_0+A_1h_n+A_2h_n^2+\cdots+A_Jh_n^J+R_J(h_n)=A(h_n), \tag{6.30}$$

$$R_J(h_n)=O(h_n^{J+1}),$$

其中系数 $A_0, A_1,\cdots, A_J$ 与 h_n 无关．由 (6.30) 看出，ζ_0 可形式地看成当 $h=0$ 时对应于 $A(h)$ 的值． 因而为估计 ζ_0，一种很自然的想法是作 $A(h)$ 的一个近似 $\widetilde{A}(h)$，然后取 $\widetilde{A}(0)$ 作为 ζ_0 的一个估计．最简单的是对于 N' 中的二个数 n_1, n_2，作线性插值函数

$$\widetilde{A}(h)=\zeta_{n_1}+\frac{\zeta_{n_2}-\zeta_{n_1}}{h_{n_2}-h_{n_1}}(h-h_{n_1}), \tag{6.31}$$

且作为 $A(h)$ 的近似．令 $h=0$，得到 ζ_0 的近似

$$\xi_0 = \widetilde{A}(0) = \frac{\zeta_{n_1}h_{n_2} - \zeta_{n_2}h_{n_1}}{h_{n_2} - h_{n_1}}. \tag{6.32}$$

这是最简单的 Richardson 过程．由 (6.30) 可知

$$\xi_0 = \zeta_0 + O(h_{n_1}h_{n_2}),$$

误差阶比 ζ_{n_1} 和 ζ_{n_2} 均要小．

这一节介绍 Richardson 外插方法的一些理论基础，文中沿用上一节的记号．

令 ζ_n 是问题(6.4)的解，它是 E_n 空间中的元．为了形式地处理，我们必须把计算得到的有限个 ζ_n 归化到共同的空间中去．为此，引进下面的构造．考虑应用到原始问题 $\mathscr{B} = \{E, E^0, F\}$ 的离散化方法 $\mathscr{M} = \{E_n, E_n^0, \Delta_n, \Delta_n^0, \varphi_n\}_{n\in N'}$．对于给定的 Banach 空间 $\hat{E}$ 和具有 $\|\hat{\Delta}\| = 1$ 的线性映象 $\hat{\Delta}: E \to \hat{E}$，如果存在一个无限集 $\bar{N} \subset N'$ 和线性映象序列 $\pi_n: E_n \to \hat{E}$，$n \in \bar{N}$，使有

$$\pi_n \Delta_n = \hat{\Delta} \text{ 和 } \lim_{n\to\infty} \|\pi_n\| = 1, \tag{6.33}$$

则 $\mathscr{M}$ 称作 $\hat{\Delta}$ 可归化的，并称 π_n 是对应的归化映象．

例 6.7 考虑例 6.1

a) 考虑任意固定的有理数 $\hat{t} = \dfrac{p}{q} \in [0, 1]$，并取 $\hat{E} = R$，$\hat{\Delta} y = y(\hat{t})$，$y \in E$．于是对于 $\bar{N} = \{iq \mid i \in N\}$ 和 $\pi_n \eta_n = \eta_n\left(\dfrac{p}{q}\right)$，$\eta_n \in E_n$，$n \in \bar{N}$，$\mathscr{M}$ 是 $\hat{\Delta}$ 可归化的．

b) 取某个固定的 $\bar{n} \in N$ 和 $\hat{E} = E_{\bar{n}}$，$\hat{\Delta} = \Delta_{\bar{n}}$，则对

$$\bar{N} = \{i\bar{n} \mid i \in N\},$$

$\mathscr{M}$ 是 $\hat{\Delta}$ 可归化的，π_n 是 G_n 上的函数在 $G_{\bar{n}}$ 上的限制，$n \in \bar{N}$．

如果 $\mathscr{M}$ 是 $\hat{\Delta}$ 可归化的和如果 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散误差具有到阶 J 的渐近展开式

$$\varepsilon_n = \Delta_n \left[\sum_{j=p}^{J} \frac{1}{n^j} e_j\right] + O(n^{-(J+1)}), (n \to \infty). \tag{6.34}$$

于是由定义 6.3 和 (6.33)，有

$$\pi_n\zeta_n = \hat{\Delta}z + \hat{\Delta}\sum_{j=p}^{J}\frac{1}{n^j}\ e_j + O(n^{-(J+1)}),\ n\in\bar{N}. \qquad (6.35)$$

这表示 $\mathscr{M}$ 的 $n\in\bar{N}$ 的解 ζ_n 的归化 $\pi_n\zeta_n\in\hat{E}$ 的序列 $\{\pi_n\zeta_n\}_{n\in\bar{N}}$ 在固定空间 $\hat{E}$ 中具有渐近展开．注意，对于 Richardson 外插需要的是渐近展开(6.35)，而不是展开式(6.34)．展开式(6.34)仅是展开式(6.35)的存在性的充分条件，而不是必要条件，于是可以提出下面的定义．

定义 6.9 如果应用到 $\mathscr{B}$ 的离散化方法 $\mathscr{M}$ 是 $\hat{\Delta}$ 可归化的，并且 $\mathscr{M}$ 得到的解序列 $\{\zeta_n\}$ 满足

$$\pi_n\zeta_n = \hat{\Delta}z + \sum_{j=p}^{J}\frac{1}{n^j}\hat{e}_j + O(n^{-(J+1)}),\ n\in\bar{N}, \qquad (6.36)$$

其中 $\hat{e}_j\in\hat{E}$, $j=p, p+1,\cdots, J$ 与 n 无关，则称 $\mathscr{M}$ 对 $\mathscr{B}$ 的整体离散误差具有 $\hat{\Delta}$ 归化的到阶 J 的渐近展开．

对于某个域 I 和空间 X，令 $C_r\subset(I\to X)$, $r\in N$, 是一族由 I 到 X 的函数，使得给定离散点 $n_\rho\in I$ 和 $\chi_\rho\in X$, $\rho=0, 1,\cdots, r$，恰好存在一个函数 $\chi\in C_r$ 使有

$$\chi(n_\rho) = \chi_\rho,\ \rho=0, 1,\cdots, r, \qquad (6.37)$$

这样的函数族的序列 $\{C_r\}_{r\in N}$ 称作由 I 到 X 的插值函数类．

多项式集合构成由 R 到 R 的插值函数类．在这种情形，族 C_r 是次数不超过 r 的多项式类．

插值函数类 C 称作是线性的，如果每个 C_r 均是线性空间，否则称它是非线性的．

现在考虑 $\mathscr{B}$ 的离散化方法 $\mathscr{M}$．假设它是 $\hat{\Delta}$ 可归化的，并且它对 $\mathscr{B}$ 的离散误差具有一个到 J 阶的 $\hat{\Delta}$ 归化渐近展开(6.36)．再假定有一个由 N 到 $\hat{E}$ 的插值函数类 C，使 C 中的所有函数在无穷远处具有有限的极限．如果我们确定了 $\mathscr{M}$ 的解序列 $\{\zeta_n\}$ 的 $r+1$ 个元 ζ_{n_ρ}, $n_\rho\in\bar{N}$, $\rho=0, 1,\cdots, r$，且元 $\pi_{n_\rho}\zeta_{n_\rho}\in\hat{E}$ 确定唯一的函数 $\chi\in C_r$，使有

$$\chi(n_\rho) = \pi_{n_\rho}\zeta_{n_\rho},\ \rho=0, 1,\cdots, r.$$

Richardson 外插或"外插到极限"是去确定极限值

$$\chi(\infty) = \lim_{n\to\infty}\chi(n) \in \hat{E}.$$

并且将其作为 $\hat{\Delta}z$ 的近似值. 这是这一节开始所描述的过程的形式描述.

为了上述过程对于满足 (6.36) 的值 $\pi_n\zeta_n$ 是合理的,我们必须应用一个插值函数类 C, 使得它以 n^{-1} 的幂的渐近展开式具有

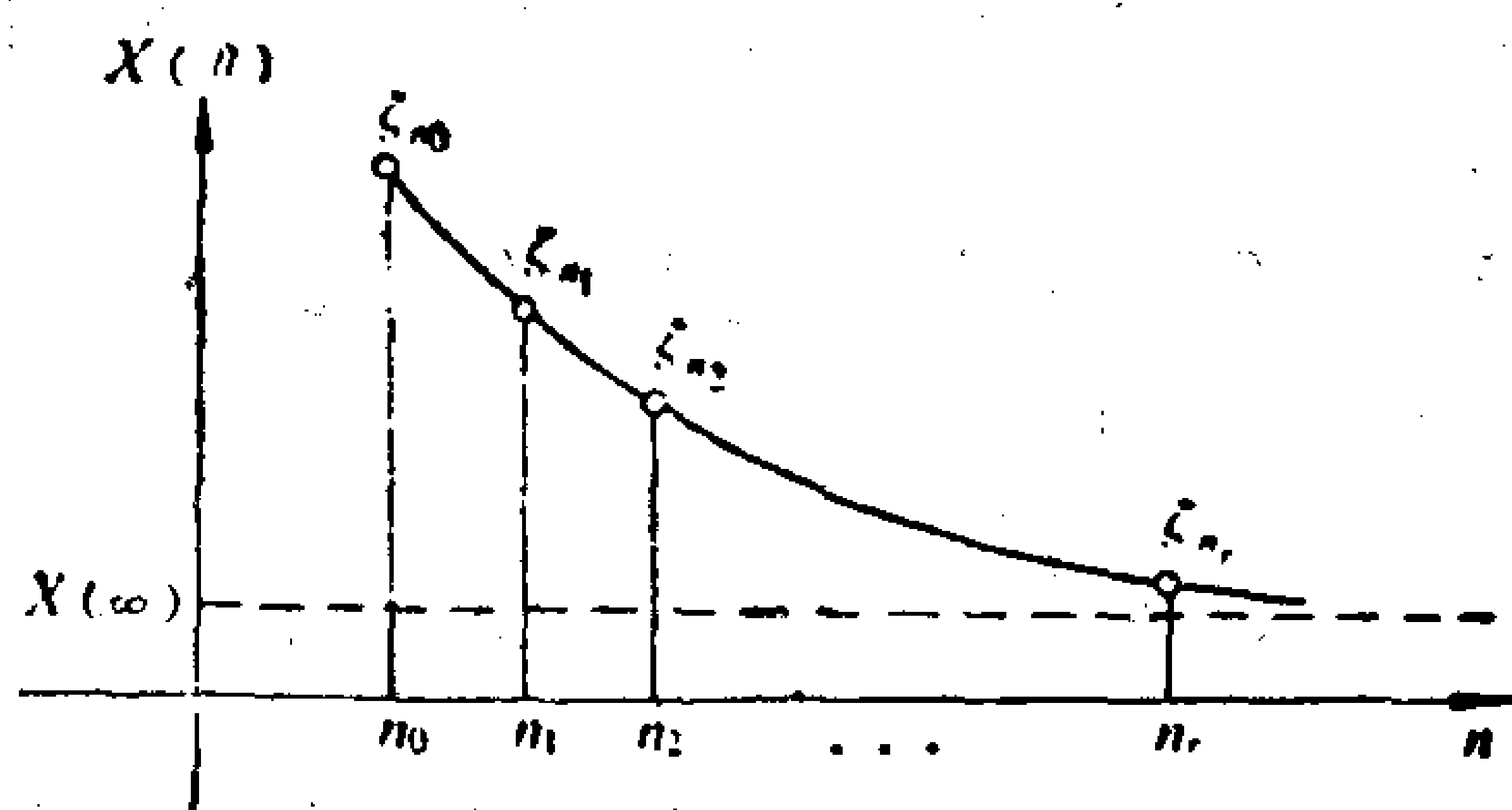


图 6.1 Richardson 外插

与(6.36) 同样的结构,即对每个 $\chi\in C$,有形式

$$\chi(n) = x_\infty + \sum_{j=p}^{J}\frac{1}{n^j}x_j + O(n^{-(J+1)}),\ x_\infty,\ x_j\in \chi.$$

并且如果 (6.36) 只含 n^{-1} 的偶次幂,则 $x\in C$ 也应该是 n^{-1} 的偶函数.

用 T_0^ρ, $n_\rho\in \bar{N}$ 表示任意的但固定的 $\pi_{n_\rho}\zeta_{n_\rho}$ 的值.由 (6.36),有

$$T_0^\rho = \bar{z} + \sum_{j=p}^{J}\frac{1}{n_\rho^j}\bar{e}_j + R_J(n_\rho), \tag{6.38}$$

其中 $\bar{z}$ 和 $\bar{e}_j$ 分别表示 $\hat{\Delta}z$ 和 $\hat{e}_j$, 当 $n\to\infty$ 时, $R_J(n)=O(n^{-(J+1)})$.

对于满足 (6.38) 的,并且有 $n_{\rho+1}>n_\rho$ 的给定的序列 $\{T_0^\rho\}$, $\rho=0, 1,\cdots$和由 $[0, 1]$ 到R的固定的插值函数类 $C=\{C_r\}$, 我们用 $\chi_r^i\in C_r$ 表示有

$$\chi_r^i\left(\frac{1}{n_\rho}\right) = T_0^\rho,\ \rho = i, i+1,\cdots, i+r$$

的函数. 当 $r\geqslant 1$ 时令

$$T_r^i = \chi_r^i(0),$$

我们得到外插表

$$\begin{array}{cccc} T_0^0 & & & \\ & T_1^0 & \ddots & \\ T_0^1 & & & T_r^0 \\ \vdots & \vdots & \cdot^{\cdot^{\cdot}} & \\ & T_1^{r-1} & & \vdots \\ T_0^r & & & \\ & \vdots & & \\ \vdots & & & \end{array}$$

下面称这个表为 T 表。

在线性外插的情形，即每个 C_r 均是线性空间，T_r^i 是 T_0^ρ，$\rho = i, i+1, \cdots, i+r$ 的线性组合

$$T_r^i = \sum_{\rho=i}^{i+r} \gamma_{r\rho}^i T_0^\rho, \tag{6.39}$$

其中系数 $\gamma_{r\rho}^i$ 依赖于插值函数类 C 和序列 $\{n_\rho\}$. 对于这种情形可以得到下面的定理.

定理 6.4 假定 $\lim\limits_{\rho\to\infty} T_0^\rho = \bar{z}$，如果对所有 i 和 r，$\gamma_{r\rho}^i$ 满足

$$\sum_{\rho=i}^{i+r} \gamma_{r\rho}^i = 1, \tag{6.40}$$

于是对每个固定的 $r > 0$，有

$$\lim_{i\to\infty} T_r^i = \bar{z}. \tag{6.41}$$

另外，如果对所有的 i 和 r 一致地有

$$\sum_{\rho=i}^{i+r} |\gamma_{r\rho}^i| \leqslant c < \infty, \ c \text{ 是某正数}, \tag{6.42}$$

且对每个固定的 i 和固定的 $\rho \geqslant i$，有

$$\lim_{r\to\infty} \gamma_{r\rho}^i = 0, \tag{6.43}$$

则对每个固定的 $i \geqslant 0$，有

$$\lim_{r\to\infty} T_r^i = \bar{z}. \tag{6.44}$$

证明 (6.41) 立即由 (6.39) 和 (6.40) 推得. 极限 (6.44) 由

Steinhaus-Toeplitz 定理推得(见关肇直：泛函分析讲义，第二章，pp159—160)。

例 6.8　令 C 是多项式类。由插值多项式的 Lagrange 表达式，有

$$\gamma_{r\rho}^{i} = \prod_{\substack{\rho'=i \\ \rho'\neq\rho}}^{i+r} \frac{-\dfrac{1}{n_{\rho'}}}{\dfrac{1}{n_\rho} - \dfrac{1}{n_{\rho'}}} = \prod_{\substack{\rho'=i \\ \rho'\neq\rho}}^{i+r} \frac{1}{1 - \dfrac{n_{\rho'}}{n_\rho}}, \tag{6.45}$$

在这种情形，条件 (6.42)，(6.43) 成立的必要充分条件是存在 $\delta > 0$，使对所有 $\rho \in N$ 有

$$\frac{n_{\rho+1}}{n_\rho} \geqslant 1 + \delta.$$

对于有限的 i 和 r，我们考虑 T_r^i 对 $\bar{z}$ 的逼近性质。

定理 6.5　令 T_0^ρ 具有到阶 $p+r$ 的渐近展开式 (6.38)，并且 $p \geqslant 1$ 的值已知，如果 C 取成为具有性质

$$\chi'(0) = \cdots = \chi^{(p-1)}(0) = 0 \tag{6.46}$$

的多项式类，则

$$T_r^i = \bar{z} + \left(\frac{1}{n_{i+r}^p} \prod_{\rho=i}^{i+r-1} \frac{1}{n_\rho}\right) K_{pr}(n_i, \cdots, n_{i+r}) \bar{e}_{p+r} + \sum_{\rho=i}^{i+r} \gamma_{r\rho}^i R_{p+r}(n_\rho), \tag{6.47}$$

其中 K_{pr} 是关于它的 $r+1$ 个变量的阶为零的齐性函数(即它仅依赖于它们的比值)。

特别 $K_{1r} \equiv (-1)^r$ 与 n_ρ 无关。所以对于 $p=1$，有

$$T_r^i = \bar{z} + (-1)^r \left(\prod_{\rho=i}^{i+r} \frac{1}{n_\rho}\right) \bar{e}_{r+1} + \sum_{\rho=i}^{i+r} \gamma_{r\rho}^i R_{r+1}(n_\rho). \tag{6.48}$$

证明　由假定，从 (6.38)，当 $J = p + r$ 时有

$$T_0^\rho = \left[\bar{z} + \sum_{i=p}^{p+r-1} \frac{1}{n_\rho^i} \bar{e}_i\right] + \frac{1}{n_\rho^{p+r}} \bar{e}_{p+r} + R_{p+r}(n_\rho). \tag{6.49}$$

由插值的线性性质，我们可以分别对这三项插值。据假设 (6.46)，

C_{p+r-1} 中的多项式有形状

$$\chi(x) = c_0 + c_p x^p + \cdots + c_{p+r-1} x^{p+r-1}, \tag{6.50}$$

(6.49) 的第一个括弧中的值是多项式 (6.50) 在 $x = \frac{1}{n_\rho}$ 处的值，$\rho = i, i+1, \cdots, i+r$，所以 $c_0 = \bar{z}$, $c_j = \bar{e}_j$, $j = p, p+1, \cdots, p+r-1$. 所以插值得到精确的值 $\chi(0) = \bar{z}$.

对于项 $\frac{1}{n_\rho^{p+r}} \bar{e}_{p+r}$ 的插值，必须寻找形式为 (6.50) 的多项式，使得有

$$\chi(x_\rho) = x_\rho^{p+r}, \quad x_\rho = \frac{1}{n_\rho}, \quad \rho = i, i+1, \cdots, i+r.$$

这就导出线性方程组

$$\begin{pmatrix} x_i^p & \cdots & x_i^{p+r-1} & 1 \\ \vdots & & \vdots & \vdots \\ x_{i+r}^p & \cdots & x_{i+r}^{p+r-1} & 1 \end{pmatrix} \begin{pmatrix} c_p \\ \vdots \\ c_{p+r-1} \\ c_0 \end{pmatrix} = \begin{pmatrix} x_i^{p+r} \\ \vdots \\ x_{i+r}^{p+r} \end{pmatrix},$$

应用 Cramer 法则，将 $\chi(0) = c_0$ 表成二个行列式之商. 将分母的行列式按它的最后一列展开，并应用 Vandermonde 行列式展开的公式，给出

$$\begin{aligned} \frac{1}{c_0} &= \sum_{\rho=i}^{i+r} \frac{1}{x_\rho^p \prod\limits_{\substack{\rho'=i \\ \rho' \neq \rho}}^{i+r} (x_\rho - x_{\rho'})} \\ &= \left[\frac{1}{x_{i+r}^p} \prod_{\rho=i}^{i+r-1} \frac{1}{x_\rho} \right] \sum_{\rho=i}^{i+r} \left[\left(\frac{x_{i+r}}{x_\rho} \right)^{p-1} \prod_{\substack{\rho'=i \\ \rho' \neq \rho}}^{i+r} \frac{x_{\rho'}}{x_\rho - x_{\rho'}} \right] \\ &= \left(n_{i+r}^p \prod_{\rho=i}^{i+r-1} n_\rho \right) [K_{pr}(n_i, \cdots, n_{i+r})]^{-1}, \end{aligned} \tag{6.51}$$

其中 $K_{pr}(n_i, \cdots, n_{i+r})$ 为

$$K_{pr}(n_i, n_{i+1}, \cdots, n_{i+r}) = \left\{ \sum_{\rho=i}^{i+r} \left[\left(\frac{x_{i+r}}{x_\rho} \right)^{p-1} \right. \right.$$

$$\cdot \prod_{\substack{\rho'=i\\ \rho'\neq\rho}}^{i+r} \frac{x_{\rho'}}{x_\rho - x_{\rho'}}\Big]\Big\}^{-1} \tag{6.52}$$

K_{pr} 的零阶齐次性是显然的.

对于 $p=1$,有

$$K_{1r} = \left[\sum_{\rho=i}^{i+r} \prod_{\substack{\rho'=i\\ \rho'\neq\rho}}^{i+r} \frac{\frac{1}{n_{\rho'}}}{\frac{1}{n_\rho} - \frac{1}{n_{\rho'}}}\right]^{-1} = \left[(-1)^r \sum_{\rho=i}^{i+r} \gamma^i_{r\rho}\right]^{-1} = (-1)^r \tag{6.53}$$

证完.

注意，只当估计式(6.38)成立时，估计(6.47)和(6.48)成立. 另外，如果 $\gamma^i_{r\rho}$ 满足(6.42)，则对于(6.47)的右边的第三项有估计式

$$\left|\sum_{\rho=i}^{i+r} \gamma^i_{r\rho} R_{p+r}(n_\rho)\right| \leqslant A \max_{\rho=i,i+1,\cdots,i+r} |R_{p+r}(n_\rho)|. \tag{6.54}$$

推论 6.2 在定理 6.5 的条件下，如果渐近展开式(6.38)是 n^{-1} 的偶次幂，$p=2$ 和 C 是偶多项式类. 于是

$$T^i_r = \bar{z} + (-1)^r \left(\prod_{\rho=i}^{i+r} \frac{1}{n^2_\rho}\right) \bar{e}_{2r+2} + \sum_{\rho=i}^{i+r} \bar{\gamma}^i_{r\rho} R_{2r+2}(n_\rho), \tag{6.55}$$

其中 $\bar{\gamma}^i_{r\rho}$ 是(6.39)式中用偶多项式插值的系数.

证明 在定理 6.5 中取 $p=2$,并用 n^2_ρ 代替 n_ρ 即可.

通常 n_ρ 取成基本参数 $\bar{n}$ 的倍数. 下面的推论给出 $\bar{n}\to\infty$ 时 T^i_r 对 $\bar{z}$ 的近似程度.

推论 6.3 在定理 6.5 的条件下,如果

$$n_\rho = \beta_\rho \bar{n},\ \rho = 0, 1, \cdots, r,\ \beta_\rho\ 固定$$

和如果对应的 $\gamma^0_{r\rho}$ 对 $\bar{n}\in N$ 一致地满足(6.42),则有

$$T^0_r = \bar{z} + O(\bar{n}^{-(p+r)}),\ \bar{n}\to\infty.$$

如果在推论 6.2 的情形,即 T^ρ_0 可展成 n^{-1} 的偶次幂,于是有

$$T^0_r = \bar{z} + O(\bar{n}^{-(2r+2)}),\ \bar{n}\to\infty,$$

证明 这个推论由估计式(6.47),(6.55)和

$$R_{p+r}(n) = O(n^{-(p+r+1)})$$

推得.

如果渐近展开式(6.38)可以展开到任意阶,且当 i 增大时,$\bar{e}_i$ 是一致有界的,以及对所用的特殊序列 $\{n_\rho\}$,K_{pr} 也是一致有界的(对于 $p=1$,这是平凡满足的),于是由(6.47)推得

$$\frac{T_{r+1}^i - \bar{z}}{T_r^i - \bar{z}} = O\left(\left(\frac{n_{i+r}}{n_{i+r+1}}\right)^{p-1} \frac{1}{n_{i+r+1}}\right),\text{当 } r \text{ 增加时,}$$

这表示沿着外插表的对角线外插是超线性收敛的.

在多项式外插的情形中,下面二种情形用外插表构造的算法是特别简单的.

(a) $p=1$,或者 $p=2$ 和展开式(6.38)是 n 的偶次幂$\left(\text{一般是}\left(\frac{1}{n_\rho}\right)^p\text{的幂次}\right)$. 在这种情形,$C$ 由所有多项式组成,或者为所有的偶多项式组成. 这时若 T_0^ρ, $\rho=0,1,\cdots$ 已计算好,则由递推式

$$T_i^\rho = T_{i-1}^{\rho+1} + \frac{1}{\left(\dfrac{n_{i+\rho}}{n_\rho}\right)^p - 1}(T_{i-1}^{\rho+1} - T_{i-1}^\rho) \tag{6.56}$$

构造 T 表.

(b) 若 $n_\rho = b^\rho n_0$, $b>1$,由(6.56)得

$$T_i^\rho = T_{i-1}^{\rho+1} + \frac{1}{b^{ip}-1}(T_{i-1}^{\rho+1} - T_{i-1}^\rho), \tag{6.57}$$

上式还可由渐近展开式(6.38)推广到

$$T_0^\rho = \bar{z} + \sum_{i=1}^{r}\left(\frac{1}{n_\rho}\right)^{\beta_i}\bar{e}_i + O(n_\rho^{-B}),\ \rho = 0,1,\cdots,r$$

的情形. 这时 $0<\beta_1<\beta_2<\cdots<\beta_r<B$ 是已知的数,但它不一定是整数. 递推地应用

$$T_i^\rho = T_{i-1}^{\rho+1} + \frac{1}{b^{\beta_i}-1}(T_{i-1}^{\rho+1} - T_{i-1}^\rho) \tag{6.58}$$

得到 T_r^0 有估计式

$$T_r^0 = z + O(n_0^{-B}) \tag{6.59}$$

§ 3 利用梯形法的整体外插

由 § 1，用梯形法求解初值问题 (6.1) 时得到的数值解 y_n 具有渐近展开式

$$y_n = y(t_n, h) = y(t_n) + A_2(t_n)h^2 + A_4(t_n)h^4 + \cdots, \tag{6.60}$$

其中 $A_2(t), A_4(t), \cdots$ 仅是 t 的函数，不依赖于步长 h 和数值解本身。另一方面，由 Dahlquist 的结果，梯形法是线性多步方法中精度阶最高，并且误差常数最小的 A 稳定方法. 由这两个理由，使用梯形法作为 Richardson 外插的基本数值方法是合理的.

利用梯形法进行 Richardson 外插可以有二种方式. 一种方式称作整体外插. 其计算过程可以这样来描述. 设求解的区间是 [0, 1]. 选定一个基本步长 h (一般 h 能等分区间)，先用梯形法以步长 h 从 0 计算到 1. 接着以步长 $\frac{h}{2}$ 从 0 计算到 1. 类似地，用梯形法分别以步长 $\frac{1}{2^i}h, i = 2, \cdots, r$ 重复地从 0 计算到 1. 根据上面的计算结果，在所需要的相同的格点上进行 Richardson 外插，得到更精确的值. 这种计算过程显然能保持方法的 A 稳定性. 这是因为在使用梯形法进行计算时，计算过程是 A 稳定的，而 Richardson 外插得到的值在后面的计算中未被利用. 因此 Richardson 外插中的误差是不会传递到下面的量，而梯形法计算的误差也不会通过 Richardson 外插传递到后面的格点上. 因而不影响稳定性.

另外一种使用方式称作局部外插，它的计算过程是这样的. 先用梯形法以 y_0 为初值分别用步长 $h, \frac{1}{2}h, \cdots, \frac{1}{2^r}h$ 从 0 积分 (6.1)，得到 $t = h$ 时的数值解 $y(t, h), y\left(t, \frac{1}{2}h\right), \cdots, y\left(t,\right.$

$\frac{1}{2^r}h)$, $y\left(t, \frac{1}{2^i}h\right)$ 表示用步长 $\frac{h}{2^i}$ 算得的数值解. 利用这些值进行 Richardson 外插得到精度阶更高的值,记这个值为 y_1^R. 用 y_1^R 作为 $t=h$ 点上的初值重复上面的过程,可得到 $t=2h$ 时的值 y_2^R. 这样往前推进直到 $t=1$ 为止.

局部外插与整体外插的区别在于整体外插可以看成是以不同的步长在整个区间上分别积分完成后, 进行 Richardson 外插, 而局部外插是每向前推进一个基本步长 h, 立即进行 Richardson 外插;利用这个外插值作为下一步的初值再向前推进.

从直观上讲,似乎局部外插比整体外插要精确,进行实际的数值试验也证明了这一点. 但是局部外插的误差会传递下去,使得整个计算过程失去梯形法所具有的 A 稳定性. 这一点可以这样来证明: 将梯形法的局部外插应用到试验方程

$$y' = \lambda y,$$

则由 $t=nh$ 点计算 $t=(n+1)h$ 点上解的近似值时, T 表的第一列有表达式

$$T_0^\rho = \left[\frac{1+(h\lambda)/2^{\rho+1}}{1-(h\lambda)/2^{\rho+1}}\right]^{2^\rho} y_n,\ \rho=0, 1, 2, \cdots, \tag{6.61}$$

其中 y_n 是前一步的局部外插得到的 $y(nh)$ 的近似值,即当前一步的积分初值. 由递推关系式 (6.57), T 表的其它列按关系式

$$T_i^\rho = T_{i-1}^{\rho+1} + \frac{1}{4^i-1}(T_{i-1}^{\rho+1} - T_{i-1}^\rho),\ \rho = 0, 1, 2, \cdots \tag{6.62}$$

递推. 因此 T_i^ρ 可以写成 T 表的第一列的线性组合的形式

$$T_i^\rho = \sum_{k=0}^{i} c_{i,i-k} T_0^{\rho+k} \tag{6.63}$$

其中系数 $c_{i,i-k}$ 满足递推关系式

$$c_{i,i-k} = \frac{4^i c_{i-1,i-k} - c_{i-1,i-1-k}}{4^i-1},\ c_{i-1,i} = c_{i-1,-1} = 0,\ c_{0,0} = 1$$

$i=1, 2, 3, \cdots, k=0, 1, \cdots, i$. 结合 (6.61) 和 (6.63), 得到

$$T_i^\rho = \sum_{k=0}^{i} c_{i,i-k}\left[\frac{1+(h\lambda)/2^{\rho+k+1}}{1-(h\lambda)/2^{\rho+k+1}}\right]^{2^{\rho+k}} y_n. \tag{6.64}$$

若将外插值 T_i^ρ 作为 y_{n+1}，则 y_{n+1} 可表成

$$y_{n+1} = \beta(h, \lambda, i, \rho) y_n, \tag{6.65}$$

其中

$$\beta(h, \lambda, i, \rho) = \sum_{k=0}^{i} c_{i,i-k}\left[\frac{1+(h\lambda)/2^{\rho+k+1}}{1-(h\lambda)/2^{\rho+k+1}}\right]^{2^{\rho+k}}. \tag{6.66}$$

$\beta(h, \lambda, i, \rho)$ 也就是递推式 (6.65) 的特征根，由它来确定局部外插过程的稳定性质.

考虑 $i=1, \rho=0$ 的情形，即考虑 T 表的第二列第一个元素，得

$$\beta(h, \lambda, 1, 0) = \frac{4}{3}\left[\frac{1+h\lambda/4}{1-h\lambda/4}\right]^2 - \frac{1}{3}\left[\frac{1+h\lambda/2}{1-h\lambda/2}\right]$$

$$= \frac{96-18(h\lambda)^2-5(h\lambda)^3}{96-96(h\lambda)+30(h\lambda)^2-3(h\lambda)^3}. \tag{6.67}$$

对于实数值 $h\lambda$，直接验证，具有 $|\beta(h, \lambda, 1, 0)| < 1$ 的区域为 $-25.86 < h\lambda < 0$ 和 $1.856 < h\lambda < 1.886$. 因此，对于 $i=1, \rho=0$ 局部外插过程就失去了梯形法的 A 稳定性. 由 (6.67) 还容易看出，当 $h\lambda \to \infty$ 时具有极限

$$\lim_{h\lambda\to\infty}\beta(h, \lambda, 1, 0) = \frac{5}{3},$$

由上面的特例，可以作下面的推理，对应于其它的 i 和 ρ 值的局部外插过程也不一定是 A 稳定的. [7] 对几个 i 和 ρ 画出了局部外插的稳定区域，证实了这个推理. 但是由 [7] 中的稳定区域可以看出，对于中等程度刚性比的问题 (6.1)，梯形法的局部外插公式是可以用的，并且计算精度比整体外插要稍高一点.

由上面的讨论可以看到，为了应用 Richardson 外插方法仍保持 A 稳定性，就必须采用整体外插.

利用梯形法的整体外插方法可以这样来进行.

1. 用梯形法以基本步长 h 从 0 到积分区间的终端积分初值问

题(6.1)，将得到的结果记成 $T^0_{0,n}$，$n=1,2,\cdots$，附加的足标表示在点 nh 处的值.

2. 从 0 到终端分别以步长 $h_\rho=h/2^\rho$，$\rho=1,2,\cdots,r$ 用梯形法积分常微分方程初值问题(6.1)，并将对应于格点 $t_n=nh$ 上的结果记成 $T^\rho_{0,n}$，$n=1,2,\cdots$，$\rho=1,2,\cdots,r$. 于是对应于每个格点 $t_n=nh$，构造了 T 表的第一列.

3. 对每一个 $t_n=nh$，$n=1,2,\cdots$，根据由 1，2. 构成的 T 表的第一列，按次序(省略足标 n)

$$\begin{array}{l} T_i^\rho \searrow \\ T_i^{\rho+1}\rightarrow T_{i+1}^\rho \end{array}$$

及公式(6.62)进行外插，构成 T 表的其余诸列，并且比较构成的列的相邻元素 T_i^ρ 和 T_ρ^{i+1}，或者比较 T 表上的主对角线上的相邻元素 T_i^0 和 T_{i+1}^0，若它们之间的差小于预给的精度，则外插结束，并将得到的值记为 y_n，作为精确解 $y(t_n)$ 的近似值.

4. 如果对某个 n，T 表的主对角线上最后相邻几个元按预给的精度相差太大，则令 $r=r+1$，计算 T 表的第 $r+1$ 行，再进行比较. 这种过程重复进行，直到满足精度为止.

在实际计算时，为了节省计算机的贮存量，可以不必象 1. 中所说的将所有的 n 的 T 表计算好再进行外插，而可以在点 t_n 处进行外插后，再用各个步长的相应值作为初值计算下一个点 t_{n+1} 的 T 表，进行外插.

为了保证数值积分结果具有渐近展开式(6.60)，在每一步必须保证递推式

$$y_{s+1}=y_s+\frac{\bar{h}}{2}\left(f(t_s,y_s)+f(t_{s+1},y_{s+1})\right) \tag{6.68}$$

是精确的. 这里所谓精确是指误差小于渐近展开式(6.60)中的截断误差. 特别由于(6.68)是隐式的，这就要求精确地求解非线性方程组. 通常采用 Newton 方法，若将上步的 y_s 作为迭代的初值，或者将前几个格点上的 y_s 进行外插得到 y_{s+1} 的初值，在很少几次迭代后就可达到所需要的精度.

§4 平滑过程

虽然梯形法是 A 稳定的，但由于当 $h\lambda \to -\infty$ 时，其特征值有

$$\frac{1 + h\lambda/2}{1 - h\lambda/2} \to -1.$$

梯形法不是刚性 A 稳定的. 对应于大的负特征值的解分量，若它们在计算开始时不是小到可忽略时，由梯形法积分得到的这些分量的近似值具有衰减很慢随步数振荡的性质，使得总的数值解在计算开始时是不精确的. 这表示如果步长选得不很小，即使梯形法是 A 稳定的，计算结果也是不精确的. 这种性质将影响 Richardson 外插的精度. 为了克服这个缺点，一种方法是在对应于大的特征值的分量未衰减到可忽略的程度时，采用小步长进行数值积分. 只要这些分量衰减到可忽略的程度，即达到解的非刚性部分精度的允许范围之内，由于 A 稳定性，可以选用大的步长进行积分. 这时方法的 A 稳定性将能保持这些刚性分量在数值误差的允许范围内，从而保持整个解的精度.

另外一种克服这种数值振荡的方法是在计算的最初几步应用类似于 Milne 和 Reynolds [86]，[87] 中使用的平滑过程. 这里的平滑表示如果已经计算出 y_{k-1}，y_k 和 y_{k+1}，则用量

$$\hat{y}_k = (y_{k-1} + 2y_k + y_{k+1})/4 \tag{6.69}$$

来代换 y_k，并且将 $\hat{y}_k$ 作为后面计算的起始值. 当问题的暂态部分具有很强的非线性时，这种方法可能是特别有效的. 下面对梯形法详细地分析这种过程.

1. 一个步长，平滑一次的算法

对于问题 (6.1) 用步长 h 进行积分，并要求在 $t = h_0$ 处进行平滑，其中 h_0 固定，$h_0 = ph$，p 是整数. 对于 $s = 0, 1, 2, \cdots, p$，由

$$y_{s+1} - y_s = \frac{h}{2}(f(t_s, y_s) + f(t_{s+1}, y_{s+1})) \tag{6.70}$$

求出 y_{s+1},计算

$$\hat{y}_p = (y_{p-1} + 2y_p + y_{p+1})/4, \tag{6.71}$$

并用它来代替 y_p,即对 y_p 赋以 $\hat{y}_p$ 的值. 接着,对 $s = p, p+1, \cdots$ 继续用公式 (6.70) 解出 y_{s+1}. 这样求得的 y_p, y_{p+1} 与由 (6.70) 求得的不平滑值将是不同的.

2. 一个步长,平滑 m 次的算法

类似于 1,用步长 h 积分,并要求在点 μh_0, $\mu = 1, 2, \cdots, m$ 处平滑,其中 h_0 固定,$h_0 = ph$, p 是整数. 对 $s = 0, 1, \cdots, p$. 由 (6.70) 求得 y_{s+1}, 由 (6.71) 算出 $\hat{y}_p$, 并用它代替 $\hat{y}_p$. 接着由 (6.70) 对 $s = p, p+1, \cdots, 2p$ 计算 y_{s+1},再算出

$$\hat{y}_{2p} = (y_{2p-1} + 2y_{2p} + y_{2p+1})/4, \tag{6.72}$$

并用它代替 y_{2p},继续上面的过程,直到计算出 $\hat{y}_{mp}$,并用 $\hat{y}_{mp}$ 代替 y_{mp}. 继续用 (6.70) 将问题积分到终点.

3. q 个步长,平滑 m 次的算法

类似于 1, 但使用 q 个步长 $h_0, h_0/2, \cdots, h_0/2^{q-1}$, 并在 m 个点 $t = \mu h_0$, $\mu = 1, 2, \cdots, m$ 处进行平滑. 对于每个步长 h,用 2 中的算法计算, 将得到的解记为 $y(t, h)$. 下面我们将证明 $y(t, h)$ 仍具有仅含 h 的偶次幂的渐近展开式,即有

$$y(t, h) = y(t) + \sum_{v=1}^{M} A_v(t)h^{2v} + O(h^{2M+1}), \tag{6.73}$$

利用这个展开式,在 $t = \mu h_0$ 的每个点上,令 $T_0^\rho = y(t, h_0/2^\rho)$, 则由递推式 (6.62) 构成 T 表,得到解的更精确的近似值.

为了证明公式 (6.73),我们首先指出在本章 §1 中实际上已证明的定理(见例 6.3 和定理 6.2).

定理 6.6 设初值问题

$$y' = f(t, y), \ y(0) = \eta_0$$

的右函数 $f(t, y)$ 在 t, y 的乘积空间中的某个区域中是 $2M$ 次连续可微的,如果用梯形法求解这个问题,并且假设其起始值可表

成

$$y_0 = \eta_0 + \sum_{v=1}^{M} q_v h^{2v} + O(h^{2M+1}), \qquad (6.74)$$

将计算结果记成序列 $\{y_k\}_{k=0,1,\cdots}$, 则有渐近展开式

$$y_k = y(t_k) + \sum_{v=1}^{M} c_v(t_k) h^{2v} + O(h^{2M+1}) \qquad (6.75)$$

其中 $t_k = kh$, $c_v(t)$ 是 t 的 $2M+1-2v$ 次连续可微的函数.

利用这个定理,可以证明展开式 (6.73) 是成立的.

定理 6.7 如果 $f(t, y)$ 对 t, y 是 $2M+1$ 次连续可微的,则利用 2 中定义的算法,而得到的解对 h 具有渐近展开式 (6.73).

证明 在区间 $[0, h_0 + h]$ 上以步长 h 用梯形法求解,其起始值为 y_0, 用 $y(t)$ 表示精确解,用 $y(t, h)$ 表示用步长 h 得到的解. 由定理 6.6,对所有的 v,令 $q_v = 0$,则在这个区间中我们有展开式

$$y(t, h) = y(t) + \sum_{v=1}^{M} c_v(t) h^{2v} + O(h^{2M+1}), \qquad (6.76)$$

其中 $y(t) \in C^{2M+2}$, $c_v(t) \in C^{2M+1-2v}$.

平滑 (6.71) 表示

$$\hat{y}(h_0, h) = [y(h_0 - h, h) + 2y(h_0, h) + y(h_0 + h, h)]/4, \qquad (6.77)$$

按 (6.76),得到

$$\begin{aligned}\hat{y}(h_0, h) = \frac{1}{4}\Big\{ & y(h_0 - h) + \sum_{v=1}^{M} c_v(h_0 - h) h^{2v} + O(h^{2M+1}) \\ & + 2\Big[y(h_0) + \sum_{v=1}^{M} c_v(h_0) h^{2v} + O(h^{2M+1})\Big] \\ & + y(h_0 + h) + \sum_{v=1}^{M} c_v(h_0 + h) h^{2v} + O(h^{2M+1})\Big\}. \qquad (6.78)\end{aligned}$$

由于 $y(t)$, 和 $c_v(t)$ 的可微性,可以将 (6.78) 的右边项在点 $t = h_0$ 展成 Taylor 级数,经整理后,可以将 (6.78) 表成

$$\hat{y}(h_0, h) = y(h_0) + \sum_{v=1}^{M} g_v(h_0)h^{2v} + O(h^{2M+1}). \quad (6.79)$$

现在接着用梯形法求微分方程在区间 $[h_0, 2h_0+h]$ 上的解，而在点 $t = h_0$ 上的起始值 $y(h_0)$ 用计算得到的量 $\hat{y}(h_0, h)$ 代替。由定理 6.6 和 (6.79)，在这区间上计算得到的解将具有仅含步长 h 的偶次幂的渐近展开式。类似地，将上述的讨论继续推广到整个积分区间，定理证毕。

下面讨论平滑过程对积分刚性方程的效果。将 2 中讨论的算法应用到初值问题

$$y' = \lambda y \quad y_0 = y(0) = 1, \ \mathrm{Re}\lambda < 0, \quad (6.80)$$

用梯形法以步长 h 积分，得到

$$y_n = \frac{1 + h\lambda/2}{1 - h\lambda/2} y_{n-1} = \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^n y_0.$$

设我们在 $t_p = ph$ 处平滑，则有

$$\begin{aligned}\hat{y}_p &= [y_{p-1} + 2y_p + y_{p+1}]/4 \\ &= \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{p-1} \left[1 + 2\,\frac{1 + h\lambda/2}{1 - h\lambda/2} + \left(\frac{1 + h\lambda/2}{1 - h\lambda/2}\right)^2\right] \\ &= \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{p-1} \frac{1}{(1 - h\lambda/2)^2} y_0.\end{aligned}$$

若在 $t_p, t_{p+1}, \cdots, t_{p+m-1}$ 处均进行平滑，得到

$$\hat{y}_{p+1} = \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{p-1} \frac{1}{(1 - h\lambda/2)^4} y_0,$$

$$\vdots$$

$$\hat{y}_{p+m-1} = \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{p-1} \frac{1}{(1 - h\lambda/2)^{2m}} y_0.$$

一般我们可以推得下面的公式。若按 2 中的算法，在 $t_n = nh$ 处和在它前面的点上共进行 m 次平滑，则有

$$y_n = \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{n-m} \frac{1}{(1 - h\lambda/2)^{2m}} y_0. \quad (6.81)$$

由 (6.81) 看出，当 $h\lambda$ 的值相当大时，通过若干次平滑，可以使用

梯形法求解方程(6.80),所引起的数值振荡很快地衰减下来.

关于平滑过程的误差分析,[73] 引进了误差函数

$$e_m(n, -h\lambda/2) = e^{h\lambda n} - \left[\frac{1 + h\lambda/2}{1 - h\lambda/2}\right]^{n-m} \frac{1}{(1 - h\lambda/2)^{2m}}.$$

它是步数 n, 平滑数 m 和 $-h\lambda/2$ 的函数. [73] 对这个函数进行了比较详细的讨论. 从讨论的结果看出,如果 $|-h\lambda/2|$ 很大,则平滑的次数愈多,误差愈小. 而对于 $|-h\lambda/2| < 1$ 的分量,当平滑次数大,而步数 n 较小时,误差将变大,当 n 增大时,误差将与不平滑时相似. 当 $h\lambda/2$ 为实数时还将稍有改进. 根据这种定性的性质可以看出,对于刚性方程组,仅在大的刚性分量出现暂态时使用平滑过程比较合适. 在出现暂态时,可立即使用若干步平滑过程,然后用一般的梯形法进行计算. 这样一方面可以使方法引起的数值振荡迅速衰减,另一方面又不影响非刚性分量的精度.

§5 用内插法求中间点上高精度近似值

用 §2 中描述的 Richardson 外插方法求解常微分方程初值问题时,首先选定一个基本步长 h_0,分别用 $h_0, h_0/2, \cdots, h_0/2^r$ 作为步长进行数值积分. 然后在以 h_0 为间隔的格点上进行外插,求出这些格点上解的更精确的近似值. 但是若需要在这些格点的中间点上的值,按常规需要将 h_0 缩小, 从而加大计算量. Lindberg[74] 根据 Dahlquist 的思想推导一个算法, 利用内插法得到中间点上精度高的近似值.

设用某种数值方法以步长 h 求解初值问题 (6.1) 得到的数值解为 $y(t, h)$,并且 $y(t, h)$ 具有渐近展开式

$$y(t, h) = y(t) + \sum_{\nu=1}^{N} c_\nu(t) h^{\nu p} + O(h^{(N+1)p}), \qquad (6.82)$$

其中 $c_\nu(t)$ 是定义在 $t \geqslant 0$ 上的充分光滑的函数, p 是固定的正整数,通常为 1 或 2. 为方便起见,假定 (6.1) 中的右函数 $f(t, y)$ 是无限可微的.

选定基本步长 h_0，令 $h_s = h_0/2^s$. 定义格点集 $G_s = \{t \mid t = ih_s, i = 0, 1, \cdots, 2^s\}$, $s = 0, 1, \cdots, q$. 令 $h = h_q$ 和 $y_0(t, h_s) = y(t, h_s)$, 按 (6.57) 类似的推导,取

$$\alpha_{k-1} = 1/(2^{kp} - 1). \tag{6.83}$$

由递推式

$$y_k(t, h) = y_{k-1}(t, h) + \alpha_{k-1}[y_{k-1}(t, h) - y_{k-1}(t, 2h)] \tag{6.84}$$

作外插,有估计式

$$y_k(t, h) - y(t) = O(h^{(k+1)p}). \tag{6.85}$$

由 (6.82) 推得 $y_k(t, h)$ 具有渐近展开式

$$y_k(t, h) = y(t) + \sum_{\nu=k+1}^{N} K_{k\nu} c_\nu(t) h^{\nu p} + O(h^{(N+1)p}), \tag{6.86}$$

其中

$$K_{k\nu} = \prod_{\gamma=1}^{N} \frac{2^{\gamma p} - 2^{\nu p}}{2^{\nu\gamma} - 1},$$

$y_k(t, h)$ 是定义在格点集 G_{q-k} 上的，特别，$y_q(t, h)$ 是确定在 G_0 上的. 我们的目的是想在点集 G_q 上确定一个函数 $\hat{y}_q(t)$,使有估计

$$\hat{y}_q(t) = y(t) + O(h^{(q+1)p}).$$

这里及下面假定 $q \leqslant N$.

由 (6.86),对于 $t \in G_0$,得到

$$\begin{aligned} y_q(t, h) - y_{k-1}(t, h) &= \sum_{\nu=q+1}^{N} K_{q,\nu} c_\nu(t) h^{\nu p} \\ &\quad - \sum_{\nu=k}^{N} K_{k-1,\nu} c_\nu(t) h^{\nu p} + O(h^{(N+1)p}) \\ &= -\sum_{\nu=k}^{q} K_{k-1,\nu} c_\nu(t) h^{\nu p} \\ &\quad + \sum_{\nu=q+1}^{N} (K_{q,\nu} - K_{k-1,\nu}) c_\nu(t) h^{\nu p} + O(h^{(N+1)p}) \end{aligned}$$

$$= -\sum_{\nu=k}^{q} K_{k-1,\nu}c_\nu(t)h^{\nu p} + O(h^{(q+1)p}). \tag{6.87}$$

在区间 $I = [0, h_0]$ 上定义函数

$$z_{k-1}(t) = -\sum_{\nu=k}^{q} K_{k-1,\nu}c_\nu(t)h^{\nu p}, \ k = q, q-1, \cdots, 1. \tag{6.88}$$

由 (6.85),(6.87),(6.88),对于 $t \in G_0$,有

$$\begin{aligned} y_{k-1}(t, h) + z_{k-1}(t) &= y_q(t, h) + O(h^{(q+1)p}) \\ &= y(t) + O(h^{(q+1)p}). \end{aligned} \tag{6.89}$$

由此,为确定 $\hat{y}_q(t)$,只要能确定 $z_{k-1}(t)$ 的近似 $\hat{z}_{k-1}(t)$,然后由它及 $y_{k-1}(t, h)$ 来构造 $\hat{y}_q(t)$. 即对 $t \in G_{q-k+1}$,令

$$\hat{y}_q(t) = y_{k-1}(t, h) + \hat{z}_{k-1}(t), \tag{6.90}$$

在 G_0 上,令 $\hat{y}_q(t) = y_q(t, h)$,所以 $\hat{y}_q(t)$ 在 G_0 上,即在 I 的二个点 α_1, α_2 上是已知的. $y_{q-1}(t, h)$ 在 G_1 上,即在 I 的三个点 $\alpha_1, \alpha_2, \alpha_3$ 上是已知的. 于是在 G_0 上,由 (6.89),(6.90),若取

$$\hat{z}_{q-1}(\alpha_i) = \hat{y}_q(\alpha_i) - y_{q-1}(\alpha_i, h), \tag{6.91}$$

则有

$$\begin{aligned} z_{q-1}(\alpha_i) &= \hat{y}_q(\alpha_i) - y_{q-1}(\alpha_i, h) + O(h^{(q+1)p}) \\ &= \hat{z}_{q-1}(\alpha_i) + O(h^{(q+1)p}), \ i = 1, 2. \end{aligned} \tag{6.92}$$

用线性内插来构造 $\hat{z}_{q-1}(\alpha_3)$

$$\hat{z}_{q-1}(\alpha_3) = \hat{z}_{q-1}(\alpha_1) + \frac{\alpha_3 - \alpha_1}{\alpha_1 - \alpha_2}(\hat{z}_{q-1}(\alpha_1) - \hat{z}_{q-1}(\alpha_2)), \tag{6.93}$$

如果对 $i = 1, 2$ 认为有 $\hat{z}_{q-1}(\alpha_i) = z_q(\alpha_i)$,则 $\hat{z}_{q-1}(\alpha_3) - z_{q-1}(\alpha_3)$ 将等于内插的绝对误差. 内插的相对误差为 $O(h^2)$,而 $z_{q-1}(t)$ 和它的导数的量级为 $O(h^{qp})$.因此,内插的绝对误差是 $O(h^{qp+2})$. 由 (6.92),我们得到

$$\hat{z}_{q-1}(\alpha_3) = z_{q-1}(\alpha_3) + O(h^{\beta_q}), \tag{6.94}$$

其中 $\beta_q = \min\{(q+1)p, pq+2\}$,即

$$\beta_q = \begin{cases} q+1 & \text{如果 } p = 1, \\ pq+2 & \text{如果 } p \geqslant 2. \end{cases} \tag{6.95}$$

由 (6.90)，得到 $\hat{y}_q(\alpha_3)$，并且由 (6.86)，(6.90)，(6.94) 有

$$\hat{y}_q(\alpha_3) - y(\alpha_3) = O(h^{\beta_q}).$$

下面对愈来愈多的点递推地定义 $\hat{y}_q$，令 $k<q$，假定 $\hat{y}_q(t)$ 在格点集 G_{q-k} 上已定义，即在 I 中的 $2^{q-k}+1$ 个点 $\alpha_1, \alpha_2, \cdots, \alpha_{2^{q-k}+1}$ 个点上有定义，并且有估计

$$\hat{y}_q(t) - y(t) = O(h^{\beta_{k+1}}).$$

于是类似 (6.90)，对于 $t \in G_{q-k}$，$\hat{z}_{k-1}(t)$ 是已知的，并且有

$$\hat{z}_{k-1}(t) - z_{k-1}(t) = \hat{y}_q(t) - y_{k-1}(t, h) - (y(t) - y_{k-1}(t,h)) + O(h^{(q+1)p}) = \hat{y}_q(t) - y(t) + O(h^{(q+1)p}) = O(h^{\beta_{k+1}}).$$

利用分段 m_k 次多项式内插，$\hat{z}_{k-1}(t)$ 可以对所有 $t \in I$ 定义。对于 $t \in I$ 内插的相对误差是 $O(h^{m_k+1})$。由 (6.88)，有

$$z_{k-1}(t) = O(h^{kp}),$$

因此内插的绝对误差是 $O(h^{kp+m_k+1})$。考虑到在 G_{q-k} 上 $\hat{z}_{k-1}(t) - z_{k-1}(t)$ 的阶，对于 $t \in G_{q-k+1}$，有

$$\hat{z}_{k-1}(t) - z_{k-1}(t) = O(h^{\beta_k}),$$

其中

$$\beta_k = \min\{\beta_{k+1}, kp + m_k + 1\}, \quad k-1<q,$$

β_q 由 (6.95) 确定.

由于 G_{q-k} 中有 $2^{q-k}+1$ 个点．因此最大可能的 m_k 是 2^{q-k}．虽然在前面指出要求 $\hat{y}_q(t)$ 中的误差不能超过 $O(h^{(q+1)p})$，但是选取 m_k 有 $m_k + kp + 1 > (q+1)p$ 是不一定好的，特别对于等距内插可能出现 Runge 现象。由于这一点，按照

$$m_k = \begin{cases} 2^{q-k}, & \text{如果 } (q-k)p + p - 1 > 2^{q-k}, \\ (q-k)p + p - 1, & \text{其它}. \end{cases}$$

选取 m_k，进行内插，使得在 G_{q-k+1} 的格点上得到 $\hat{y}_q(t)$，并且具有估计

$$\hat{y}_q(t) - y(t) = O(h^{\beta_k}).$$

由归纳假定，这式对于 $k=q$ 成立，因此，对任意 k 的值 $0 \leqslant k < q$ 均成立．这样递推地构造，对于 G_q 中的格点，构造了 $\hat{y}_q(t)$，

并且有估计

$$\hat{y}_q(t) = y(t) + O(h^\beta), \tag{6.96}$$

其中

$$\beta = \min_{k=1,2,\cdots,q} \{(q+1)p,\ kp + 2^{q-k} + 1\}, \tag{6.97}$$

对于具体的 t 值，对应的 β 值可能比 (6.97) 给出的要大。

例 9 将上述算法应用到梯形法。取 $q = 3$，则有 $p = 2$，$\beta = 7$，$\alpha_\nu = 1/(4^\nu - 1)$。可以按下面的计算步骤进行。

1. 用梯形法计算

(a) $y_0(t_k, h)$，$h = h_0/2^3$，$t_k = kh$，$k = 1, 2, \cdots, 8$。

(b) $y_0(t_k, 2h)$，$k = 2, 4, 6, 8$。

(c) $y_0(t_k, 4h)$，$k = 4, 8$。

(d) $y_0(t_8, 8h)$。

这些量均有误差 $O(h^2)$。将它们排成表 6.1

表 6.1

步长	h	$2h$	$4h$	$8h$
t_0	$y_0(t_0, h)$	$y_0(t_0, 2h)$	$y_0(t_0, 4h)$	$y_0(t_0, 8h)$
t_1	$y_0(t_1, h)$			
t_2	$y_0(t_2, h)$	$y_0(t_2, 2h)$		
t_3	$y_0(t_3, h)$			
t_4	$y_0(t_4, h)$	$y_0(t_4, 2h)$	$y_0(t_4, 4h)$	
t_5	$y_0(t_5, h)$			
t_6	$y_0(t_6, h)$	$y_0(t_6, 2h)$		
t_7	$y_0(t_7, h)$			
t_8	$y_0(t_8, h)$	$y_0(t_8, 2h)$	$y_0(t_8, 4h)$	$y_0(t_8, 8h)$

2. 对于每个 t_ν，逐次计算 $\{y_k(t_\nu, h)\}_{k=1}^3$。例如，对于 $\nu = 4$，我们计算

$$y_1(t_4, h) = y_0(t_4, h) + \frac{1}{3}[y_0(t_4, h) - y_0(t_4, 2h)]$$

$$y_2(t_4, h) = y_1(t_4, h) + \frac{1}{15}[y_1(t_4, h) - y_1(t_4, 2h)]$$

于是得到表 6.2. 类似于由 (6.82) 得 (6.86),有

$$y_k(t, h) = y(t) + \sum_{\nu=k+1}^{\infty} K_{k,\nu} c_\nu(t) h^{2\nu},$$

定义

$$\hat{y}_3(t_8) = y_3(t_8, h),$$
$$\hat{y}_3(t_0) = y_3(t_0, h),$$

表 6.2

t_0	$y_0(t_0, h)$	$y_1(t_0, h)$	$y_2(t_0, h)$	$y_3(t_0, h)$
t_1	$y_0(t_1, h)$			
t_2	$y_0(t_2, h)$	$y_1(t_2, h)$		
t_3	$y_0(t_3, h)$			
t_4	$y_0(t_4, h)$	$y_1(t_4, h)$	$y_2(t_4, h)$	
t_5	$y_0(t_5, h)$			
t_6	$y_0(t_6, h)$	$y_1(t_6, h)$		
t_7	$y_0(t_7, h)$			
t_8	$y_0(t_8, h)$	$y_1(t_8, h)$	$y_2(t_8, h)$	$y_3(t_8, h)$
误　差	$O(h^2)$	$O(h^4)$	$O(h^6)$	$O(h^8)$

3. 由表 6.2 和公式 (6.90)

$$\hat{z}_2(t_i) = \hat{y}_3(t_i) - y_2(t_i, h) + O(h^8), \quad i = 0, 8,$$

再由线性插值给出 $\hat{z}_2(t_4)$,使有

$$\hat{z}_2(t_4) - z_2(t_4) = O(h^2 h^6) = O(h^8).$$

因此有

$$\hat{y}_3(t_4) = y_2(t_4, h) + \hat{z}_2(t_4) + O(h^8).$$

对于 t_i, $i = 0, 4, 8$,令

$$\hat{z}_1(t_i) = \hat{y}_3(t_i) - y_1(t_i, h) + O(h^8),$$

而用在这三个点上的二次插值算出 $\hat{z}_1(t_2)$ 和 $\hat{z}_1(t_6)$, 并且 $\hat{z}_1(t_2) - z_1(t_2)$ 和 $\hat{z}_1(t_6) - z_1(t_6)$ 均是 $O(h^3 h^4) = O(h^7)$ 的. 所以有

$$\hat{y}_3(t_i) = y_1(t_i, h) + \hat{z}_1(t_i) = y(t_i) + O(h^7) \quad i = 2, 6,$$

由此,我们有

$$\hat{z}_0(t_i) = \hat{y}_3(t_i) - y_0(t_i, h) = z_0(t_i) + \begin{cases} O(h^7) & i = 2, 6, \\ O(h^8) & i = 0, 4, 8. \end{cases}$$

由这些量应用四次插值给出 $\hat{z}_0(t_i)$, $i = 1, 3, 5, 7$,且有

$$\hat{z}_0(t_i) - z_0(t_i) = O(h^5 \cdot h^2) = O(h^7).$$

于是有

$$\hat{y}_3(t_i) = y_0(t_i, h) + \hat{z}_0(t_i) = y(t_i) + O(h^7),$$
$$i = 1, 3, 5, 7.$$

对于这样构造的 $\hat{y}_3(t_i)$ 有

$$\hat{y}_3(t_i) = y(t_i) + \begin{cases} O(h^7) & i = 1, 2, 3, 5, 6, 7, \\ O(h^8) & i = 0, 4, 8. \end{cases}$$

§6 应用平滑和外插的隐式中点方法

隐式中点方法是 A 稳定的,具有梯形法的类似性质,并且应用起来比梯形法还要方便一点.因此也经常用来求解刚性方程组.

为简化记号,考虑自守系统

$$y' = f(y), \quad y(0) = y_0. \tag{6.98}$$

应用隐式中点公式

$$\begin{aligned} &y(0, h) = y_0 \\ &y(t_i + h, h) = y(t_i, h) + hf([y(t_i + h, h) \\ &\qquad + y(t_i, h)]/2), \\ &t_i = ih \end{aligned} \tag{6.99}$$

求解 (6.98),其中 h 为步长, $y(t, h)$ 是计算得到的 (6.98) 的解的近似。由例 6.6 和 (6.99) 得到的 $y(t, h)$,有渐近展开式

$$y(t_n, h) = y(t_n) + h^2 d_1(t_n) + h^4 d_2(t_n) + W_n(h), \tag{6.100}$$

其中 $y(t)$ 是 (6.98) 的精确解. 由定理 6.1 构造展开式的系数,可得

$$d_1'(t) - J(t)d_1(t) = \frac{1}{12} y'''(t) - \frac{1}{8} J'(t)y'(t), \tag{6.101}$$

$$d_2'(t) - J(t)d_2(t) = -y^{(V)}(t)/120 + d_1'''(t)/12$$

$$+ \frac{1}{2} f_{yy}''(y(t))(d_1(t) + y''(t)/8)^2$$

$$- \frac{1}{384}(y^{(V)}(t) - J(t)y^{(IV)}(t))$$

$$- \frac{1}{8}(d_1'''(t) - J(t)d_1''(t) - y^{(V)}(t)/12), \quad (6.102)$$

$$W_{n+1}(h) - W_n(h) = hJ_{n+\frac{1}{2}}(W_{n+1}(h) + W_n(h))/2 + O(h^7), \quad (6.103)$$

这里 $J(t) = \left(\frac{\partial f}{\partial y}\right)_{y(t)}$,

$$J_{n+\frac{1}{2}} = \int_0^1 \left(f_y'(y(t_{n+\frac{1}{2}}) + \theta\left[\frac{1}{2}(y(t_n + h, h) + y(t_n, h)) - y(t_{n+\frac{1}{2}})\right]\right) d\theta.$$

展开式 (6.100) 构成了应用隐式中点方法的 Richardson 外插方法求解刚性方程的基础. 但是在 (6.100) 中的 $W_n(h)$ 中对刚性方程组可能含有振荡的分量，类似于 (6.69) 的平滑程序可用来减小振荡的振幅. 平滑可以这样进行：若已计算出 $y(t_k - h, h)$, $y(t_k, h)$ 和 $y(t_k + h, h)$，计算

$$\hat{y}(t_k, h) = [y(t_k + h, h) + 2y(t_k, h) + y(t_k - h, h)]/4. \quad (6.104)$$

对于这个 $\hat{y}(t_k, h)$，有展开式

$$\hat{y}(t_k, h) = y(t_k) + [d_1(t_k) + y''(t_k)/4]h^2 + [d_2(t_k) + d_1''(t_k)/4 + y^{(IV)}(t_k)/48]h^4 + \hat{W}_k(h) + O(h^6), \quad (6.105)$$

其中

$$\hat{W}_k(h) = [W_{k-1}(h) + 2W_k(h) + W_{k+1}(h)]/4.$$

展开式 (6.105) 中只含有步长 h 的偶次幂，应用 Richardson 外插得到比 $\hat{y}(t_k, h)$ 更精确的值. 令

$$\bar{y}(t_k, h) = \hat{y}\left(t_k, \frac{h}{2}\right) + \left[\hat{y}\left(t_k, \frac{h}{2}\right) - \hat{y}(t_k, h)\right]/3, \tag{6.106}$$

则有估计式

$$\begin{aligned}\bar{y}(t_k, h) = y(t_k) - (d_2(t_k) + d_1''(t_k)/4 \\ + y^{(IV)}(t_k)/48)h^4/4 + \overline{W}_k(h) + O(h^6),\end{aligned} \tag{6.107}$$

其中

$$\overline{W}_k(h) = \hat{W}_k\left(\frac{h}{2}\right) + \left[\hat{W}_k\left(\frac{h}{2}\right) - \hat{W}_k(h)\right]/3.$$

Lindberg[75] 叙述了利用公式(6.99)，(6.100)，(6.104)和(6.105)的一个计算机程序的方案．程序从给定的初始点积分到指定的最后时刻 T．在工作过程中，重复下面四个基本的计算步．

(*a*) 进行积分和平滑．

(*b*) 估计误差．

(*c*) 确定是否接受这一步，确定下一步所用的步长．

(*d*) 为下一步计算作准备．

现在对这四步分别作一些简单的说明

(*a*) 用给定的步长 h，应用隐式中点法则计算二个独立的解 $y(t, h)$ 和 $y\left(t, \frac{h}{2}\right)$，

$$y(t+h, h) - y(t, h) = hf([y(t+h, h) + y(t, h)]/2) \tag{6.108}$$

$$t = 0, h, 2h, \cdots,$$

$$\begin{aligned}y\left(t+\frac{h}{2}, \frac{h}{2}\right) - y\left(t, \frac{h}{2}\right) = \frac{h}{2} f\left(\left[y\left(t+\frac{h}{2}, \frac{h}{2}\right)\right.\right. \\ \left.\left. + y\left(t, \frac{h}{2}\right)\right]\Big/2\right)\end{aligned} \tag{6.109}$$

$$t = 0, \frac{h}{2}, h, \frac{3}{2}h, \cdots,$$

在点 $t_k = kh$，$k = 1, 2, \cdots$ 处作平滑．得

$$\hat{y}(t_k, h) = [y(t_k + h, h) + 2y(t_k, h) + y(t_k - h, h)]/4,$$

$$\hat{y}\left(t_k, \frac{h}{2}\right) = \left[y\left(t_k + \frac{h}{2}, \frac{h}{2}\right) + 2y\left(t_k, \frac{h}{2}\right) + y\left(t_k - \frac{h}{2}, \frac{h}{2}\right)\right] \Big/ 4.$$

由这些平滑值作一次外插，得

$$\bar{y}(t_k, h) = \left[\hat{y}\left(t_k, \frac{h}{2}\right) + \left[\hat{y}\left(t_k, \frac{h}{2}\right) - \hat{y}(t_k, h)\right] \Big/ 3.$$

这里得到的平滑值和外插值在以后的计算中将不再应用，只作为点 t_k 处的计算结果.

非线性方程组 (6.108) 和 (6.109) 用拟 Newton 方法求解. 令 $z = y(t + h, h)$, $y = y(t, h)$, 和

$$F(z) = z - y - hf([z + y]/2),$$

则方程组 (6.108) 可写成为 $F(z) = 0$，可由

$$H(z^{i+1} - z^i) = -F(z^i)$$

进行迭代求解，其中 H 是矩阵

$$\frac{\partial F}{\partial z} = I - \frac{h}{2}\left\{\frac{\partial f}{\partial \eta}\Big|_{\eta=(z+y)/2}\right\}$$

的近似. 只要收敛速度满意，若干个迭代步可以应用同样的 H，收敛速度由

$$\rho = \|z^{i+1} - z^i\| / \|z^i - z^{i-1}\|, \quad i = 1, 2, \cdots$$

来估计. 如果 $\rho \leqslant 0.2$，认为收敛速度是满意的. 如果 $\rho > 0.2$，要重新计算迭代矩阵.

用同样的方式求解 (6.109).

为了得到 $y\left(t_{k-1} + \frac{h}{2}, \frac{h}{2}\right)$ 和 $y\left(t_k, \frac{h}{2}\right)$ 的迭代解的起始值，由 $y\left(t_{k-3}, \frac{h}{2}\right)$, $y\left(t_{k-2}, \frac{h}{2}\right)$ 和 $y\left(t_{k-1}, \frac{h}{2}\right)$ 作二次外插. 对于求解 $y(t_k, h)$ 的起始值，取

$$y\left(t_k, \frac{h}{2}\right) + \left(y(t_{k-1}, h) - y\left(t_{k-1}, \frac{h}{2}\right)\right).$$

(*b*) 误差估计

误差估计是根据公式 (6.107) 和确定 $d_1(\cdot)$ 和 $d_2(\cdot)$ 的微分方程来设计的. 下面假定与

$$(d_2(t_k) + d_1''(t_k)/4 + y^{(\mathrm{IV})}(t_k)/48)h^4/4$$

相比, $W_k(h)$, $W_k\left(\frac{h}{2}\right)$ 均是可忽略的.

定义

$$\tilde{y}(t_k, h) = y\left(t_k, \frac{h}{2}\right) + \left(y\left(t_k, \frac{h}{2}\right) - y(t_k, h)\right)\Big/3, \tag{6.110}$$

则

$$\tilde{y}(t_k, h) = y(t_k) - d_2(t_k)h^4/4 + \widetilde{W}_k(h) + O(h^6), \tag{6.111}$$

其中

$$\widetilde{W}_k(h) = W_k\left(\frac{h}{2}\right) + \left(W_k\left(\frac{h}{2}\right) - W_k(h)\right)\Big/3.$$

将 (6.107) 与 (6.111) 相比较,由平滑过程增加的误差项是

$$s(t_k) = -(d_1''(t_k) + y^{(\mathrm{IV})}(t_k)/12)h^4/16. \tag{6.112}$$

由于平滑得到的结果后面的计算不再应用, 整体误差中的这部分量不会传播下去, 但它依赖于点上的 $d_1''(\cdot)$ 和 $y^{(\mathrm{IV})}(\cdot)$ 的值. 平滑误差 $s(t_k)$ 的一个粗糙的估计为

$$\bar{s}(t_k) = \nabla^2\left[\hat{y}\left(t_k, \frac{h}{2}\right) - \hat{y}(t_k, h)\right]\Big/12. \tag{6.113}$$

由 (6.105),

$$\bar{s}(t_k) = (d_1''(t_k) + y^{(\mathrm{IV})}(t_k)/4)h^4/16 + O(h^6), \tag{6.114}$$

整体误差项 $d_2(t)h^4/4$ 在一步中的改变量为

$$l(t_k) = (d_2(t_k + h) - d_2(t_k))h^4/4, \tag{6.115}$$

称它为局部误差. 这里不对局部误差进行严格的估计, 而想得到一个比较易于实现的估计. [75] 在一系列简化的假定下推得一个近似的估计 $l(t_k) \approx \bar{l}(t_k)$,

$$\tilde{l}(t_k) = l^*(t_k)/3,$$

其中

$$l^*(t) = \nabla_h^3 \left[\hat{y}\left(t, \frac{h}{2}\right) - \hat{y}(t, h) \right] \Big/ 36, \qquad (6.116)$$

$$\nabla_h z(t) = z(t) - z(t-h).$$

(c) 步长的选取

步长的选取由局部误差的估计 $L_n = \|\tilde{l}(t_n)\|$ 来决定. 要求保持不等式 $L_n < \varepsilon$ 成立,其中 ε 是允许的最大局部误差. 由于改变步长所需要的工作量相当大,只当步长有大的改变时才进行. 另外,只当 $L_n < L_{n-1}$ 时,步长才允许增加.

用 h 表示原先的步长,而用 h_{New} 表示新的步长. 下面的表列出程序中进行步长控制的一种策略

L_n/ε	h_{New}
$<1/5120$	$(\varepsilon/(5L_n))^{1/n}\cdot h$
$[1/5120,\ 1/80]$	$2h$
$[1/80, 1/2]$	h
$[1/2,\ 1]$	$h/2$
>1	拒绝上一步,并令 $h_{\text{New}} = h/2$

步长改变时,应用下面的算法: 令原先的步长为 h 和 $h/2$,新的步长为 $H, H/2$. 由解 $y(t, h)$ $y\left(t, \frac{h}{2}\right)$ 外插到H和 $H/2$ 来得到光滑的数值解 $y(t, H)$ 和 $y\left(t, \frac{H}{2}\right)$ 的起始值.

§7 利用梯形公式局部外插的数值方法

直接利用梯形公式的外插方法必须进行整体外插,才能保持计算过程的A稳定性. 这样做的结果,使得计算过程中要贮存较

多的信息，另外，它还有两个缺点。其一，当 $|\lambda h|$ 的值充分大时，梯形方法本身收敛很慢，外插出的较精确的值无法利用；其二，在检验数值解是否满足精度要求时，要看相邻两次整体外插值之差的绝对值是否小于给定的容许误差值，这样就使得计算量大大增加。事实上，梯形公式的局部外插方法仍是可用的。如 [7] 所指出的，比起传统的 Runge-Kutta 等方法。局部外插法的稳定区域对于中等程度的刚性方程已够大的了，并且试验的精度比整体外插要好一点。

孙耿在 [11] 中应用梯形公式外插得到的值及用 Simpson 公式得到的值作适当的线性组合，构造了一个刚性稳定的单步四阶的算法簇。此外，这些方法在得到数值解的同时，还可以得到局部截断误差的具体数值，这就使得判别数值解是否满足精度要求变得非常简单。这节叙述 [11] 中构造的算法。

为叙述简单起见，假定 (6.1) 是单个方程。用 $y(t)$ 表示 (6.1) 的真解。用 y_n 表示 (6.1) 在格点 $t_n = nh$ 处的数值解，h 表示积分步长。

在区间 $[nh, (n+1)h]$ 上应用梯形公式

$$y_{n+1} = y_n + \frac{h}{2}[f(t_n, y_n) + f(t_{n+1}, y_{n+1})] \quad (6.117)$$

解 (6.1) 的数值解记为 $y_{n+1} = y(t_{n+1}, h)$，若将 (6.117) 中的 h 改成 $\frac{h}{2}$，连续两次使用这个公式所得到的 $(n+1)h$ 点处的数值解记为 $y\left(t_{n+1}, \frac{h}{2}\right)$，用 Richardson 外插，得

$$\tilde{y}(t_{n+1}, h) = y\left(t_{n+1}, \frac{h}{2}\right) + \left[y\left(t_{n+1}, \frac{h}{2}\right) - y(t_{n+1}, h)\right]\Big/3. \quad (6.118)$$

在这过程中，假定在 (6.117) 中求 y_{n+1} 是精确的。令 $y_{n+1} = \tilde{y}(t_{n+1}, h)$，则 y_{n+1} 也是问题 (6.1) 的数值解。

现将 $y(t)$ 代入 (6.118) 的计算过程，并在 $t_n = nh$ 处 Taylor

展开可得 (6.118) 的局部截断误差为

$$l_{n+1}(h) = \sum_{j\geq 5}\left[\frac{2^j}{j!} - \frac{(4 + 2^{j-1})}{3(j-1)!}\right]\left(\frac{h}{2}\right)^j y^{(j)}(t_n),$$

上式右端恰好就是以 $\frac{h}{2}$ 为步长的 Simpson 公式的误差表示式. 这就是说，梯形公式一次局部外插的局部截断误差(假定 y_n 精确)与以 $\frac{h}{2}$ 为步长的 Simpson 公式的局部截断误差(假定 y_n, $y_{n+\frac{1}{2}}$ 是已知的精确解)是相同的. 因此取 $y_{n+1} = \tilde{y}(t_{n+1}, h)$ 具有四阶精度，其局部截断误差的主项为 $-\frac{1}{2880}h^5y_n^{(5)}$. 在应用 Simpson 公式

$$y_{n+2} = y_n + \frac{h}{3}(y'_n + 4y'_{n+1} + y'_{n+2}) \tag{6.119}$$

求初值问题 (6.1) 的数值解时，若取 $y'_{n+1} = f(t_{n+1}, \tilde{y}(t_{n+1}, h))$，则 (6.119) 也具有四阶精度，其局部截断误差的主项为 $-\frac{1}{120}h^5y_n^{(5)}$.

利用上述的说明来构造初值问题 (6.1) 的数值解. 假定在 $t = 2nh$ 处的数值解 y_{2n} 已知，现在计算初值问题 (6.1) 在点 $t = (2n+1)h$ 和 $t = (2n+2)h$ 处的数值解 y_{2n+1}, y_{2n+2}.

1. 以 h, $2h$ 为步长，用梯形公式 (6.117) 作局部一次外插，分别得到 $\tilde{y}(t_{2n+1}, h)$, $\tilde{y}(t_{2n+2}, 2h)$. 令

$$y_{2n+1} = \tilde{y}(t_{2n+1}, h),\quad y_{2n+2,E} = \tilde{y}(t_{2n+2}, 2h).$$

并记 $y'_{2n+1} = f(t_{2n+1}, y_{2n+1})$.

2. 用 Simpson 公式

$$y_{2n+2,s} = y_{2n} + \frac{h}{3}(y'_{2n} + 4y'_{2n+1} + y'_{2n+2,s}) \tag{6.120}$$

求解 $y_{2n+2,s}$，并计算出 $y_{2n+2,E} - y_{2n+2,s}$. 在求解非线性方程 (6.120) 时，$\tilde{y}(t_{2n+2}, 2h)$ 可以作为预估值.

3. 选择适当的常数 α，构造线性组合

$$y_{2n+2}=\alpha y_{2n+2,E}+(1-\alpha)y_{2n+2,s}$$

使得构造的方法是刚性稳定的，所得到的 y_{2n+1}，y_{2n+1} 就是初值问题(6.1)分别在点 $t=(2n+1)h$，$t=(2n+2)h$ 处的数值解.

记上述构造数值解的方法为方法 I.

可以看到，在方法的构造过程中，只使用了梯形公式和 Simpson 公式，而所构造的方法可以认为是以 $2h$ 为步长的单步法，仿照[104]中的方法，可以证明这样构造的方法是收敛的. 关于方法的局部截断误差，根据本节上面的说明，易知有

$$y_{2n+1}-y(t_{2n+1})\approx-\frac{1}{2880}h^5y_{2n}^{(5)},$$

$$y_{2n+2}-y(t_{2n+2})\approx-\frac{3+\alpha}{360}h^5y_{2n}^{(5)},$$

由于在 2 中已算出 $y_{2n+2,E}-y_{2n+2,s}$，并且有

$$y_{2n+2,E}-y_{2n+2,s}\approx-\frac{1}{360}h^5y_{2n}^{(5)},$$

所以局部截断误差可以表示成

$$y_{2n+1}-y(t_{2n+1})\approx\frac{1}{8}(y_{2n+2,E}-y_{2n+2,s}),$$

$$y_{2n+2}-y(t_{2n+2})\approx(3+\alpha)(y_{2n+2,E}-y_{2n+2,s}).$$

综合算法 I 的求解过程，可以看到，每使用梯形公式(6.117)五次，用 Simpson 公式(6.120)一次，可以得到(6.1)的二个新点上的近似值. 所构造的方法是一种特殊的块单步方法.

关于方法 I 的稳定性，有下面的定理.

定理 6.8 当 α 满足不等式

$$\frac{5}{7}<\alpha<\frac{13}{14} \tag{6.121}$$

时，方法 I 看成为 $2h$ 步长的单步法是刚性稳定的.

证明 将方法 I 用于试验方程

$$y'=\lambda y,\ \mathrm{Re}\lambda<0.$$

经简单运算，得到

$$y_{2n+1}=\zeta_1(\mu)y_{2n}=\zeta_1(\mu)\cdot\zeta_2^n(\mu,\alpha)y_0,$$
$$y_{2n+2}=\zeta_2(\mu,\alpha)y_{2n}=\zeta_2^{n+1}(\mu,\alpha)y_0,\quad \mu=\lambda h.$$

其中特征根

$$\zeta_1(\mu)=\frac{96-18\mu^2-5\mu^3}{96-96\mu+30\mu^2-3\mu^3},$$

$$\zeta_2(\mu,\alpha)= \tag{6.122}$$

$$\frac{1-\frac{5}{6}\mu-\frac{25}{48}\mu^2+\frac{9}{48}\mu^3+\frac{101}{576}\mu^4-\frac{32\alpha-9}{288}\mu^5+\frac{28\alpha-23}{576}\mu^6}{1-\frac{7}{6}\mu+\frac{151}{48}\mu^2-\frac{85}{48}\mu^3+\frac{103}{192}\mu^4-\frac{4}{48}\mu^5+\frac{1}{192}\mu^6}.$$

在复平面上画出 $|\zeta_2(\mu,\alpha)|=1$ 的曲线(图 6.2 中的曲线 I 为特征根 $\left|\zeta_2\left(\mu,\frac{23}{28}\right)\right|=1$ 的曲线). $\zeta_1(\mu)$, $\zeta_2(\mu,\alpha)$ 的全部极点都在闭曲线所包围的正实轴上,另一方面,当 $\mu\to\infty$时,$|\zeta_2(\mu,\alpha)|<1$, 因此由极大模原理,在闭曲线外部恒有 $|\zeta_2(\mu,\alpha)|<1$. 按照刚性稳定性的定义, 当 α 满足 (6.121) 时, 方法 I 是刚性稳定

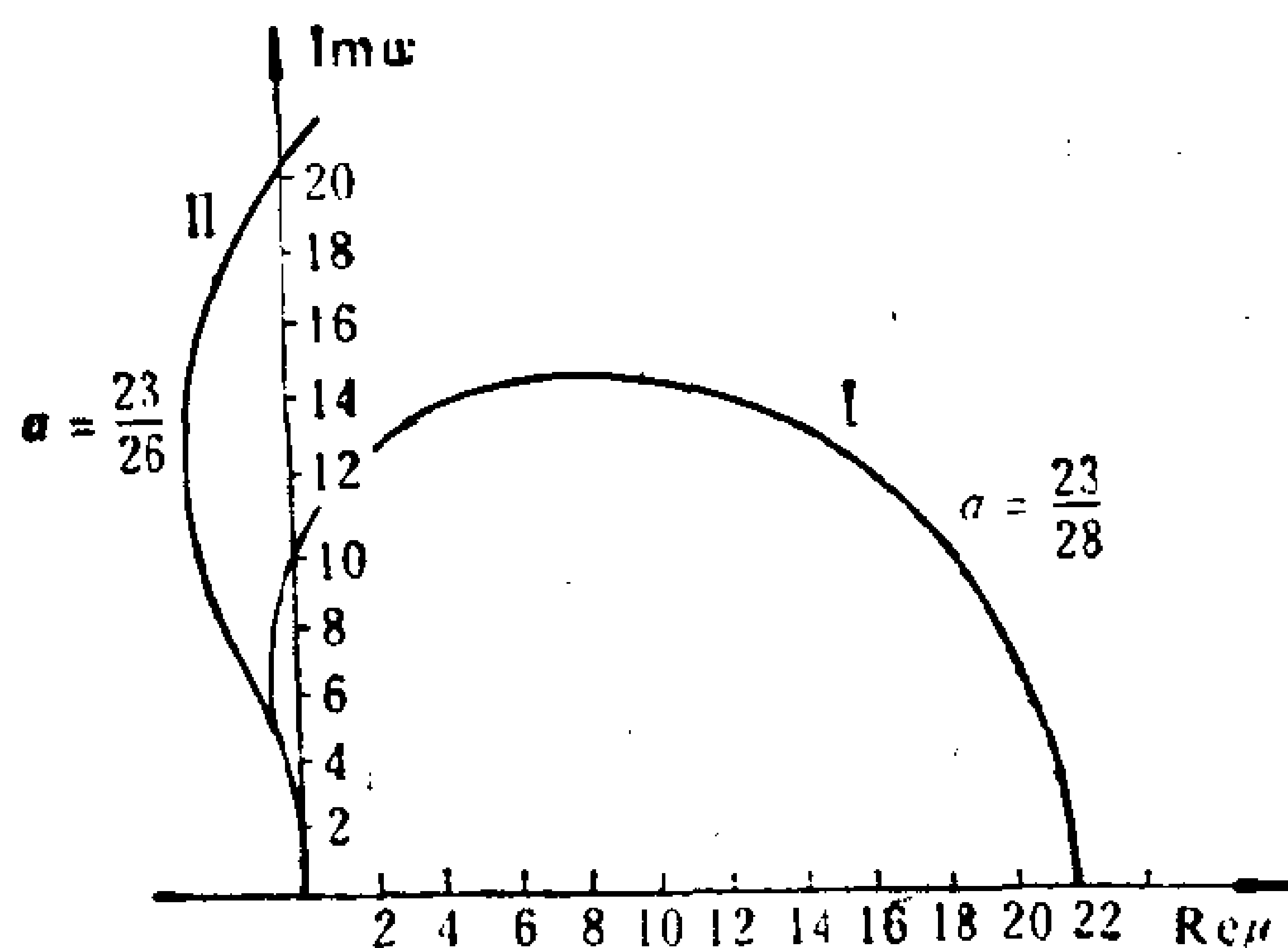


图 6.2 曲线 I 和曲线 II 的另一支是与实轴对称的. 曲线 I 的外部为方法 I 当 $\alpha=\frac{23}{28}$ 时的稳定区域,曲线 II 的左半平面为方法 II 当 $\alpha=\frac{23}{26}$ 时的稳定区域.

的.

当 $\alpha=\dfrac{23}{28}$ 时，(6.121) 仍满足，$\zeta_2(\mu,\alpha)$ 的分子 μ^6 的系数为零. 因此，对于充分大的 $|\lambda h|$ 的值，方法 I 求得的数值解有最大的衰减速度.

利用梯形公式 (6.117) 还可以构造下面的算法 II.

1. 在区间 $[2nh,(2n+2)h]$ 上以 h 为步长用梯形公式 (6.117) 作整体外插，得到 $\tilde{y}(t_{2n+1},h)$，$\tilde{y}(t_{2n+2},h)$，令 $y_{2n+1}=\tilde{y}(t_{2n+1},h)$，$y_{2n+2,E}=\tilde{y}(t_{2n+2},h)$，$y'_{2n+1}=f(t_{2n+1},y_{2n+1})$，$y_{2n+2,E}$ 的局部截断误差的主项为 $-\dfrac{1}{1440}h^5y_{2n}^{(5)}$.

第 2 步和第 3 步同方法 I. 这样得到的 y_{2n+1}, y_{2n+2} 分别是 (6.1) 在 $(2n+1)h$, $(2n+2)h$ 处的数值解. 仿照定理 6.8，可得

定理 6.9 当 α 满足

$$\frac{55}{64}<\alpha<1 \tag{6.123}$$

时，方法 II 对于以 $2h$ 为步长的单步法是刚性稳定的. 方法 II 的局部截断误差可有估计

$$y_{2n+1}-y(t_{2n+1})=-\frac{1}{24}(y_{2n+2,E}-y_{2n+2,s}),$$

$$y_{2n+2}-y(t_{2n+2})=\frac{-12+11\alpha}{11}(y_{2n+2,E}-y_{2n+2,s}).$$

对于方法 II 的 $\zeta_2(\mu,\alpha)$ 有表达式

$$\zeta_2(\mu,\alpha)=B/C,$$

其中

$$B=1-\frac{1}{3}\mu-\frac{3}{8}\mu^2+\frac{1}{48}\mu^3+\frac{161}{2304}\mu^4$$
$$-\frac{(40\alpha-58)}{2304}\mu^5+\frac{128\alpha-179}{9216}\mu^6$$

$$-\frac{26\alpha-23}{9216}\mu^7,$$

$$C=1-\frac{7}{3}\mu+\frac{55}{24}\mu^2-\frac{59}{48}\mu^3+\frac{299}{2304}\mu^4$$

$$-\frac{56}{2304}\mu^5+\frac{23}{3072}\mu^6-\frac{1}{3072}\mu^7.$$

易看出，当 $\alpha=\frac{23}{26}$ 时，条件 (6.123) 仍满足，这时 $\zeta_2\left(\mu,\frac{23}{26}\right)$ 的分子 μ^7 的系数为零。因此，对于充分大的 $|\lambda h|$ 值，方法 II 有最大的衰减速度。方法 II 的稳定区域为图 6.2 中曲线 II 所划分的左半平面。

本 章 附 注

§ 1 是根据 Stetter 的书 [106] 编写的.

§ 2 的材料主要取自 [106] 和 [71].

§ 3 是根据书 [71] 中的材料编写的.

§ 4、§ 5 和 § 6 是根据 Lindberg 的 [73],[74] 和 [75] 编写的.

§ 7 是根据孙耿的文章 [11] 编写的.

第七章　具有可变系数的线性多步方法

在前面的一些章节中，构造了一些高阶 A 稳定的线性多步方法。其中有些方法的系数中含有依赖于步长 h 或方程右函数的 Jacobi 矩阵的部分，即方法的系数是可变的。似乎正是由于这些因素，使得它们能够克服 Dahlquist 的结果对常系数线性多步方法的阶的限制。这表明构造具有变系数的线性多步方法是发展刚性方程数值求解方法的一个途径。

本章对一类变系数的线性多步方法建立一些基本结果。是按照 [70] 编写的。

§1　具有可变矩阵系数的多步方法

考虑求初值问题

$$y' = f(t, y),\quad y(t_0) = y_0 \tag{7.1}$$

的数值解的多步方法

$$\sum_{j=0}^{k}\left[a_j^{(0)}I + \sum_{s=1}^{S} a_j^{(s)}h^sQ_n^s\right]y_{n+j} = h\sum_{j=0}^{k}\left[b_j^{(0)}I + \sum_{s=1}^{S-1} b_j^{(s)}h^sQ_n^s\right]f_{n+j}, \tag{7.2}$$

其中 y 和 f 是实(或复)的 m 维向量，I 是 $m\times m$ 单位矩阵。Q_n 是 $m\times m$ 实(或复)矩阵系数，它是随 n 变化的。数系数 $a_j^{(s)}, b_j^{(s)}, j=0, 1, \cdots, k, s=0, 1, \cdots, S$ 均是常数。为方便起见，约定

$$a_k^{(0)} = 1,\quad b_j^{(-1)} \equiv b_j^{(S)} \equiv 0.$$

在实际应用时，若右函数 f 的 Jacobi 矩阵 $\dfrac{\partial f(t,y)}{\partial y}$ 在区间 $[t_n, t_{n+k}]$ 上存在，可以选取 Q_n 为负 Jacobi 矩阵 $-\partial f(t, y)/\partial y$ 在

这区间上的某种近似．但是对于下面的分析，我们只需要假定对所有的 n 有

$$
\begin{aligned}
&\text{(i)}\quad \|Q_n\| \leqslant q, \\
&\text{(ii)}\quad \left[a_k^{(0)} I + \sum_{s=1}^{S} a_k^{(s)} h^s Q_n^s\right] \text{是非奇异的，}
\end{aligned}
\tag{7.3}
$$

其中 q 是与 n 无关的常数．

令 $H = [t_0, \bar{t}]$，其中 $t_0 < \bar{t} < \infty$．K_m 是实(或复) m 维 Euclid 空间．假定 (7.1) 中的 f 是由 $H \times K_m$ 到 K_m 中的连续函数，并且对于其第二个变量 y 是一致 Lipschitz 连续的．于是初值问题 (7.1) 在 $[t_0, \bar{t}]$ 上有唯一解．用 $z(t)$ 表示这个解．在上述假定下，一定存在常数 c_1 和 c_2，使对所有 $t \in [t_0, \bar{t}]$，有

$$
\|z(t)\| \leqslant c_1, \quad \|f(t, z(t))\| \leqslant c_2. \tag{7.4}
$$

定义

$$
\omega(\delta) = \max_{\substack{|t_1 - t_2| \leqslant \delta \\ t_1, t_2 \in [t_0, \bar{t}]}} \|z'(t_1) - z'(t_2)\|, \tag{7.5}
$$

则有 $\lim\limits_{\delta \to 0} \omega(\delta) = 0$.

分别由下列式子定义线性差分算子 $L^{(s)}$，$s = 0, 1, 2, \cdots, S$ 和 $\tilde{L}$

$$
L^{(s)}[y(t); h] = \sum_{j=0}^{k} [a_j^{(s)} y(t + jh) - h b_j^{(s)} y'(t + jh)],
$$

$$
s = 0, 1, 2, \cdots, S, \tag{7.6}
$$

$$
\tilde{L}[y(t); h] = \sum_{s=0}^{S} h^s Q^s L^s [y(t); h], \tag{7.7}
$$

其中 Q 是满足条件 (7.3) 的任意 $m \times m$ 矩阵．对于可微的函数 $y(t)$，$\tilde{L}[y(t); h]$ 可以表成

$$
\begin{aligned}
\tilde{L}[y(t); h] = \tilde{c}_0^{(0)} y(t) + h[\tilde{c}_1^{(0)} y^{(1)}(t) + \tilde{c}_0^{(1)} Q y(t)] + \cdots \\
+ h^l \sum_{j=0}^{\max(l, S)} \tilde{c}_{l-j}^{(j)} Q^j y^{(l-j)}(t) + \cdots.
\end{aligned}
\tag{7.8}
$$

定义 7.1 方法 (7.2) 称作具有阶 p，如果在 (7.8) 中 $\tilde{c}_{l-j}^{(j)} = 0$，$l = 0, 1, \cdots, p$ 成立，而对 $l = p + 1$ 和某个 $j = 0, 1, \cdots,$

$\max\{l, S\}$ 有 $\tilde{c}_{l-i}^{(i)} \neq 0$. 如果方法的阶 $p \geqslant 1$, 方法称作是相容的.

这样定义的阶与矩阵 Q_n 的选取无关. 容易证明: 如果阶 $p \geqslant S$ (下面导出的方法均属于这种情形), 方法 (7.2) 具有阶 p 的充分必要条件为:

$$\sum_{j=0}^{k} a_j^{(s)} = 0,$$

$$\frac{1}{l!}\sum_{j=0}^{k} j^l a_j^{(s)} = \frac{1}{(l-1)!}\sum_{j=0}^{k} j^{l-1} b_j^{(s)}, \tag{7.9}$$

$$l = 1, 2, \cdots, p-s, \quad s = 0, 1, \cdots, S.$$

对于相容的方法,按照 Henrici [63] 可以证明下面的引理.

引理 7.1 设方程 (7.1) 满足所述的条件, 且方法 (7.2) 是相容的,则有估计式

$$\|L^{(0)}[z(t); h]\| \leqslant c^{(0)} h\omega(kh),$$

$$\|L^{(1)}[z(t); h]\| \leqslant c^{(1)} c_2 h,$$

$$\|L^{(i)}[z(t); h]\| \leqslant c^{(i)} \max(c_1, hc_2), \quad i = 2, 3, 4, \cdots, S.$$

其中 c_1, c_2 和 $\omega(\cdot)$ 分别由 (7.4) 和 (7.5) 定义,而 $c^{(i)}$ 由

$$c^{(i)} = \begin{cases} \sum_{j=0}^{k} (j|a_j^{(i)}| + |b_j^{(i)}|), & i = 0, 1, \\ \sum_{j=0}^{k} (|a_j^{(i)}| + |b_j^{(i)}|), & i = 2, 3, \cdots, S. \end{cases}$$

确定.

利用引理 7.1,立即可得估计式

$$\|\tilde{L}[z(t); h]\| \leqslant \max[c^{(0)}, c^{(1)}c_2, c^{(i)}\max(c_1, hc_2)] \cdot \left[h\omega(kh) + h^2 q + \sum_{s=2}^{S} h^s q^s\right]. \tag{7.10}$$

定义方法 (7.2) 的特征多项式为

$$\rho(\zeta) = \sum_{j=0}^{k} a_j^{(0)} \zeta^{(j)}, \tag{7.11}$$

并令 $\rho(\zeta) = 0$ 的根为 ζ_p, $p = 1, 2, \cdots, k$. 其中 ζ_1 是主根. 对

于相容的方法有 $\zeta_1 = 1$.

定义 7.2 方法 (7.2) 称作满足根条件，如果 $\rho(\zeta) = 0$ 的根均有 $|\zeta_p| \leqslant 1$, $p = 1, 2, \cdots, k$, 并且模为 1 的根是单根. 如果有 $|\zeta_p| < 1$, $p = 2, \cdots, k$,则方法称作满足严格根条件.

对方法 (7.2) 的起始值加上适当的条件，则可证明下面与常系数线性多步方法类似的定理.

定理 7.1 方法 (7.2) 为收敛的充分条件为它是相容的并且满足根条件.

与 Henrici [63] 中对常系数线性多步方法的证明类似，直接应用 [63] 的引理 3.2 到公式 (7.2), 由 (7.3) 和 (7.10) 推出这个定理.

令

$$
\begin{aligned}
&\gamma_j^{(s)} = a_j^{(s)} + b_j^{(s-1)}, \quad s = 0, 1, \cdots, S, \\
&\gamma_j^{(-1)} \equiv \gamma_j^{(s)} \equiv 0, \quad s > S, \; j = 0, 1, 2, \cdots, k.
\end{aligned}
\tag{7.12}
$$

定义 7.3 方法 (7.2) 称作满足稳定化条件，如果在根 ζ_p, $p = 2, 3, \cdots, k$ 均是不同的情形有

$$
\sum_{j=0}^{k} \gamma_j^{(s)} \zeta_p^j = 0, \quad s = 1, 2, \cdots, S, \; p = 2, 3, \cdots, k, \tag{7.13}
$$

而在 $\zeta_{p^*} = \zeta_{p^*+1} = \cdots = \zeta_{p^*+m-1}$ 的情形下，将 (7.13) 中的 $p = p^*, p^* + 1, \cdots, p^* + m - 1$ 的等式分别换成

$$
\begin{aligned}
\sum_{j=0}^{k} \gamma_j^{(s)} \zeta_{p^*}^j &= \sum_{j=1}^{k} \gamma_j^{(s)} j \zeta_{p^*}^{j-1} = \cdots \\
&= \sum_{j=m-1}^{k} \gamma_j^{(s)} j(j-1)\cdots(j-m+2) \zeta_{p^*}^{j-m+1} = 0
\end{aligned}
\tag{7.14}
$$

后，(7.13) 仍成立.

定义 7.4 方法 (7.2) 称作稳定化的，如果它满足严格根条件和稳定化条件.

定义 7.5 如果 $b_k^{(s)} = 0$, $s = 0, 1, \cdots, S$, 方法 (7.2) 称作线性隐式方法；否则称为完全隐式的.

对于线性隐式方法，每一步只须步解一个线性方程组

$$\left[I + \sum_{s=1}^{S} a_k^{(s)} h^s Q_n^s\right] y_{n+k} = g,$$

其中 g 是已知的量．而对于完全隐式方法．每一步必须求解一个非线性方程组．

可以这样来说明稳定化条件的作用．如果将方法 (7.2) 应用到试验方程

$$y' = \lambda y, \ \mathrm{Re}\lambda < 0,$$

并且将 $-Q_n$ 取为 λI，则得到一个线性常系数差分方程．这个差分方程的特征多项式可以看成是由多项式 $\rho(\zeta)$ 进行扰动得到的，扰动量的量级为 $O(h)$．称这个多项式为稳定多项式．于是稳定化条件保证稳定多项式的寄生根与多项式 $\rho(\zeta)$ 的寄生根是重合的．如果另外再假定严格的根条件，则对所有 h，这些寄生根将不引起不稳定性．另外，稳定多项式的主根是 $e^{h\lambda}$ 的有理近似，我们可以选取系数，使得这个有理近似的模小于 1，从而保证方法的 A 稳定性．

作为例子，考虑下面的稳定化二步二阶方法，并且 $S = 1$．

例 7.1

$$\left[I + \frac{1}{2} hQ_n\right] y_{n+2} - [(1+\alpha)I + hQ_n] y_{n+1}$$
$$+ \left[\alpha I + \frac{1}{2} hQ_n\right] y_n = \frac{1}{2} h[(3-\alpha) f_{n+1}$$
$$-(1+\alpha) f_n], \ -1 < \alpha < 1,$$

多项式 $\rho(\zeta)$ 的根是 1 和 α．令 $f = \lambda y$, $\mathrm{Re}\lambda < 0$ 和 $Q_n = -\lambda I$，我们得到差分方程

$$\left(1 - \frac{1}{2} h\lambda\right) y_{n+2} - \left[1 + \alpha + \frac{1}{2}(1-\alpha) h\lambda\right] y_{n+1}$$
$$+ \alpha\left(1 + \frac{1}{2} h\lambda\right) y_n = 0,$$

容易看出，这个差分方程的稳定多项式的寄生根是 α．因此方法确

实是稳定化的. 主根是 $\left(1+\frac{1}{2}h\lambda\right)\Big/\left(1-\frac{1}{2}h\lambda\right)$, 方法显然是 A 稳定的.

§2 稳定化方法的阶

这一节考虑满足稳定化条件的方法(7.2)可达到的最高阶. 先证明几个引理

引理 7.2 如果方法(7.2)的阶 $p \geqslant S$,则有

$$\sum_{j=0}^{k} \gamma_j^{(s)} = \sum_{j=0}^{k}\left(\sum_{l=1}^{s} \frac{(-1)^{l+1}}{l!} j^l \gamma_j^{(s-l)}\right),$$
$$s = 0, 1, \cdots, p. \tag{7.15}$$

证明 将(7.12)代入(7.9),立即可推出这个引理.

引理 7.3 如果方法(7.2)是稳定化的,则有

$$\gamma_j^{(s)} = \gamma_k^{(s)}(c_{j-1} - c_j) + \mu_1 c_j \sum_{l=0}^{k} \gamma_l^{(s)}$$
$$j = 0, 1, \cdots, k, \quad s = 0, 1, \cdots, S, \tag{7.16}$$

其中 c_j 由

$$\mu(\zeta) \equiv \prod_{p=2}^{k} (\zeta - \zeta_p) \equiv \sum_{p=0}^{k-1} c_p \zeta^p, \ c_k \equiv 0 \equiv c_{-1}$$

确定,而令

$$\mu_1 = 1/\mu(1).$$

证明 由定义 7.3 中的条件

$$\sum_{j=0}^{k} \gamma_j^{(s)} \zeta^j \equiv \gamma_k^{(s)}(\zeta - \sigma^{(s)})\mu(\zeta)$$
$$\equiv \gamma_k^{(s)}(\zeta - \sigma^{(s)}) \sum_{p=0}^{k-1} c_p \zeta^p, s = 0, 1, \cdots, S, \tag{7.17}$$

其中 $\sigma^{(s)}$ 是常数. 让 ζ 的同幂次的系数相等,得到

$$\gamma_j^{(s)} = \gamma_k^{(s)}(c_{j-1} - \sigma^{(s)} c_j) = \gamma_k^{(s)}(c_{j-1} - c_j)$$

$$+ c_j\gamma_k^{(s)}(1-\sigma^{(s)}),$$

$$j=0,1,\cdots,k,\quad s=0,1,\cdots,S. \tag{7.18}$$

在 (7.17) 中令 $\zeta=1$,得到

$$\sum_{j=0}^{k}\gamma_j^{(s)}=\gamma_k^{(s)}(1-\sigma^{(s)})\mu(1),\ s=0,1,\cdots,S, \tag{7.19}$$

注意到如果满足根条件,则 $\mu(1)\neq 0$. 由 (7.19) 解出 $\gamma_k^{(s)}(1-\sigma^{(s)})$,代入 (7.18),引理证得.

引理 7.4 如果方法(7.2)是稳定化的. 并且有阶 $p\geqslant S$,则

$$\mu_1\sum_{j=0}^{k}\gamma_j^{(s)}=\sum_{l=1}^{s}\frac{(-1)^{l+1}}{l!}\gamma_k^{(s-l)},\ s=0,1,2,\cdots,p. \tag{7.20}$$

证明 应用归纳法. 设结果对 $s=0,1,\cdots,m<p$ 成立. 由引理 7.2,引理 7.3 和归纳假定,

$$\sum_{j=0}^{k}\gamma_j^{(m+1)}=\sum_{j=0}^{k}\left(\sum_{l=1}^{m+2}\frac{(-1)^{l+1}}{l!}j^l\gamma_j^{(m+1-l)}\right)$$

$$=\sum_{j=0}^{k}\left[\sum_{l=1}^{m+2}\frac{(-1)^{l+1}}{l!}j^l\left\{\gamma_k^{(m+1-l)}(c_{j-1}-c_j)\right.\right.$$

$$\left.\left.+\mu_1 c_j\sum_{r=0}^{k}\gamma_r^{(m+1-l)}\right\}\right]$$

$$=\sum_{j=0}^{k}\left[\sum_{l=1}^{m+2}\frac{(-1)^{l+1}}{l!}j^l\left\{\gamma_k^{(m+1-l)}(c_{j-1}-c_j)\right.\right.$$

$$\left.\left.+c_j\sum_{t=1}^{m+2-l}\frac{(-1)^{t+1}}{t!}\gamma_k^{(m+1-l-t)}\right\}\right]$$

$$=\sum_{u=1}^{m+2}\gamma_k^{(m+1-u)}\left\{\frac{(-1)^{u+1}}{u!}\sum_{j=0}^{k}j^u(c_{j-1}-c_j)\right\}$$

$$+\sum_{l=1}^{m+2}\left\{\frac{(-1)^{l+1}}{l!}\sum_{j=0}^{k}j^l c_j\right.$$

$$\left.\cdot\sum_{t=1}^{m+2-l}\frac{(-1)^{t+1}}{t!}\gamma_k^{(m+1-l-t)}\right\},$$

将右边的第二项进行整理后，我们得到

$$\sum_{j=0}^{k} \gamma_j^{(m+1)} = \sum_{u=1}^{m+2} K^{(u)} \gamma_k^{(m+1-u)},$$

其中

$$\begin{aligned}
K^{(u)} &= \frac{(-1)^{u+1}}{u!} \sum_{j=0}^{k} j^u (c_{j-1} - c_j) \\
&\quad + \sum_{l=1}^{u-1} \left\{ \frac{(-1)^{l+1}}{l!} \frac{(-1)^{u-l+1}}{(u-l)!} \sum_{j=0}^{k} j^l c_j \right\} \\
&= \frac{(-1)^{u+1}}{u!} \sum_{j=0}^{k} \left\{ j^u (c_{j-1} - c_j) \right. \\
&\quad \left. - c_j \sum_{l=1}^{u-1} \frac{u!}{l!(u-l)!} j^l \right\} \\
&= \frac{(-1)^{u+1}}{u!} \sum_{j=0}^{k} [j^u (c_{j-1} - c_j) \\
&\quad - c_j \{ (j+1)^u - j^u - 1 \}] \\
&= \frac{(-1)^{u+1}}{u!} \sum_{j=0}^{k} c_j = \frac{(-1)^{u+1}}{u!} \frac{1}{\mu_1}.
\end{aligned}$$

对于 $s = 0$，引理显然成立．引理证完．

定理 7.2 形式为(7.2)的稳定化方法可达到的最高阶为 $2S$．

证明 由 (7.12) 和引理 7.4，推得

$$\sum_{l=1}^{s} \frac{(-1)^{l+1}}{l!} \gamma_k^{(s-l)} = 0, \quad s = S+1, S+2, \cdots, p. \tag{7.21}$$

如果阶 p 超过 $2S$，则 $S+1$ 个常数 $\gamma_k^{(s)}$，$s = 0, 1, \cdots, S$ 将满足 $S+1$ 个齐次线性方程．这显然给出 $\gamma_k^{(s)} = 0$，$s = 0, 1, \cdots, S$．与假定 $\gamma_k^{(0)} \equiv a_k^{(0)} \neq 0$ 矛盾．所以阶不能超过 $2S$．

为了构造形式为 (7.2) 的 $2S$ 阶稳定化方法，考虑方法

$$\left[\sum_{s=0}^{S} \gamma_k^{(s)} (hQ_n)^s \right] \left[\sum_{j=0}^{k-1} c_j y_{n+j+1} \right]$$

$$
-\left[\sum_{s=0}^{S}\sum_{l=0}^{s}\frac{(-1)^l}{l!}\gamma_k^{(s-l)}(hQ_n)^s\right]\left[\sum_{j=0}^{k-1}c_jy_{n+j}\right]
$$

$$
=h\sum_{j=0}^{k}\sum_{s=0}^{S-1}b_j^{(s)}(hQ_n)^s(f_{n+j}+Q_ny_{n+j}), \tag{7.22}
$$

其中 c_j 是在引理 7.3 中定义。显然这个方法具有 (7.2) 的形式。容易验证，它自动地满足稳定化条件 (7.13)。通过选取 c_j 使根 ζ_p 满足 $|\zeta_p|<1$，$p=2,3,\cdots,k$，可以使 (7.22) 为稳定化方法。事实上，由引理 7.2，引理 7.3 和引理 7.4 可以推得任何一个阶 $p\geqslant S$ 的稳定化的方法 (7.2) 均可以写成 (7.22) 的形式。

现在令 $\gamma_k^{(s)}$ 由 (7.21) 确定，其中 $p=2S$。由于 $\gamma_k^{(0)}=1$，$\gamma_k^{(s)}$，$s=0,1,\cdots S$ 是唯一确定的。由常系数线性多步方法的结果，只要 k 充分大，阶为 $2S$ 的显(隐)式常系数线性 k 步方法是存在的。可以选取 $b_j^{(s)}$，$j=0,1,\cdots,k-1$ (或 k)，$s=0,1,\cdots,S$，使得对于 $p=2S$，满足必要的阶条件 (7.9)。利用这种构造得到阶为 $2S$ 的线性隐式(或全隐式)的稳定化方法 (7.2)。定理证完。

如果根据上面定理证明，$\gamma_k^{(s)}$，$s=0,1,\cdots,S$，是唯一确定的，于是由 $p=2S$ 和 (7.21) 推得有

$$
\left[\sum_{s=0}^{S}\gamma_k^{(s)}(hQ_n)^s\right]^{-1}\left[\sum_{s=0}^{S}\sum_{l=0}^{s}\frac{(-1)^l}{l!}\gamma_k^{(s-l)}(hQ_n)^s\right]
$$

$$
=e^{-hQ_n}+O(h^{2S+1}), \tag{7.23}
$$

即 (7.23) 的左边部分是 e^{-hQ_n} 的 (S,S)Padé 近似。

若将 (7.22) 应用到方程组 $y'=Ay$，并且取 $Q_n=-A$，则 (7.22) 的右边部分为零，而得到对 y_n 的差分方程。这个差分方程的特征多项式的主根是 (7.23) 的右边部分，其余的根是 ζ_p，$p=2,3,\cdots,k$。因此，方法 (7.22) 将具有好的稳定性质。

§3 可变系数多步方法的稳定性分析

为研究形式为 (7.2) 的方法的稳定性质，考虑 m 阶常系数线性方程组

$$y' = Ay,\ y(t_0) = y_0, \tag{7.24}$$

其中 A 是 $m\times m$ 阶矩阵. 设 $\lambda_i,\ i=1,\cdots,m$ 是矩阵 A 的 m 个特征值,并且有 $\mathrm{Re}\lambda_i<0$. 于是当 $t\to\infty$ 时,(7.24) 的所有解均趋向于零.

定义 7.6 形式为 (7.2) 的方法称作 $\bar{A}$ 稳定的,如果将方法以 $Q_n=-A$ 应用到 (7.24), 对所有固定正步长 h 所得到的差分方程的解,当 $n\to\infty$ 时均趋向于零.

试验方程 (7.24) 比 Dahlquist 定义 A 稳定性所用的方程稍一般些. 在 Dahlquist 的定义中(见第一章定义 1.4) 的方程对应于选取 (7.24) 中的 A 为 $A=\lambda I$, 其中 λ 是具有 $\mathrm{Re}\lambda<0$ 的复常数.

定理 7.3 阶 $p\geqslant S$ 的形式为 (7.2) 的稳定化方法是 $\bar{A}$ 稳定的充分必要条件是系数 $\gamma_k^{(s)},\ s=0,1,\cdots,S$ 使得矩阵指数 e^{hA} 的有理近似

$$R_s(hA)=\left[\sum_{s=0}^{S}\gamma_k^{(s)}(-hA)^s\right]^{-1}$$

$$\cdot\left[\sum_{s=0}^{S}\sum_{l=0}^{s}\frac{(-1)^l}{l!}\gamma_k^{(s-l)}(-hA)^s\right] \tag{7.25}$$

对所有 $h>0$ 满足

$$\sigma(R_s(hA))<1, \tag{7.26}$$

其中 $\sigma(B)$ 是矩阵 B 的谱半径.

($R_s(hA)$ 的形式 (7.25) 保证对任何 $\gamma_k^{(s)}$ 均有

$$R_s(hA)=e^{hA}+O(h^{S+1}))$$

证明 按照下面的步骤可以得到一个直接的证明: 应用引理 7.3 和引理 7.4 形式地导出 (7.22), 然后按照定理 7.2 后面的说明得到 p 阶有理近似. 但是这种证明方式不能推广来处理试验方程为 $y'=A(t)y$ 的情形. 下面的证明易进行推广.

将形式为 (7.2) 的方法以 $Q_n=-A$ 应用到 (7.24), 得到差分方程

$$\sum_{j=0}^{k} \Phi_j y_{n+j} = 0, \tag{7.27}$$

其中

$$\Phi_j = \sum_{s=0}^{S} \gamma_j^{(s)} h^s (-A)^s. \tag{7.28}$$

定义 km 维向量 η_n 为

$$\eta_n^T = [y_{n+k-1}^T, y_{n+k-2}^T, \cdots, y_{n+1}^T, y_n^T],$$

可以将方程 (7.27) 写成单步递推的形式

$$\eta_{n+1} = C\eta_n, \tag{7.29}$$

其中

$$C = \begin{bmatrix} -\Phi_k^{-1}\Phi_{k-1} & -\Phi_k^{-1}\Phi_{k-2} & \cdots & -\Phi_k^{-1}\Phi_1 & -\Phi_k^{-1}\Phi_0 \\ I & 0 & & 0 & 0 \\ 0 & I & & 0 & 0 \\ & \cdots\cdots & \cdots\cdots & \vdots & \\ 0 & 0 & & I & 0 \end{bmatrix} \tag{7.30}$$

和 I 是 $m\times m$ 单位矩阵.

考虑特征多项式 $\rho(\zeta)$ 的根 $\zeta_p, p=2, 3, \cdots, k$ 是不同的情形,并且定义 $km\times km$ 块 Vandermonde 矩阵

$$V = \begin{bmatrix} I & \zeta_2^{k-1}I & \zeta_3^{k-1}I & \cdots & \zeta_k^{k-1}I \\ I & \zeta_2^{k-2}I & \zeta_3^{k-2}I & \cdots & \zeta_k^{k-2}I \\ & \cdots\cdots & \cdots\cdots & \cdots & \\ I & \zeta_2 I & \zeta_3 I & \cdots & \zeta_k I \\ I & I & I & \cdots & I \end{bmatrix}. \tag{7.31}$$

这个矩阵的逆具有形式(见 Gautschi [56])

$$V^{-1} = \begin{bmatrix} \mu_1 I & v_{12}I & v_{13}I & \cdots & v_{1k}I \\ \mu_2 I & v_{22}I & v_{23}I & \cdots & v_{2k}I \\ \vdots & \vdots & \vdots & & \vdots \\ \mu_k I & v_{k2}I & v_{k3}I & \cdots & v_{kk}I \end{bmatrix}, \tag{7.32}$$

其中 $\mu_r, r=1, 2, \cdots, k, v_{ij}, i=1, 2, \cdots, k, j=2, \cdots, k$ 均

是常数．特别有

$$\mu_r = \left[\prod_{\substack{j=1\\j\neq r}}^{k} (\zeta_r - \zeta_j)\right]^{-1}, \tag{7.33}$$

易看出，这里定义的 μ_1 与引理7.3中定义的 μ_1 是一致的．

由于方法满足稳定化条件，有

$$\sum_{j=0}^{k} \zeta_p^j \Phi_j = 0, \tag{7.34}$$

应用这个方程和由 $V^{-1}V \equiv I$ 所得到的恒等式，我们得到

$$V^{-1}CV = \begin{bmatrix} \Delta_1 & 0 & 0 & \cdots & 0 \\ \Delta_2 & \zeta_2 I & 0 & \cdots & 0 \\ \Delta_3 & 0 & \zeta_3 I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \\ \Delta_k & 0 & 0 & \cdots & \zeta_k I \end{bmatrix}, \tag{7.35}$$

其中

$$\begin{aligned} \Delta_1 &= I - \mu_1 \Phi_k^{-1} \sum_{j=0}^{k} (\Phi_j - \gamma_j^{(0)} I), \\ \Delta_r &= -\mu_r \Phi_k^{-1} \sum_{j=0}^{k} (\Phi_j - \gamma_j^{(0)} I) = \frac{\mu_r}{\mu_1} (\Delta_1 - I), \\ & \qquad r = 2, 3, \cdots, k. \end{aligned} \tag{7.36}$$

于是，如果令

$$u = \max\{|\zeta_r|, r = 2, 3, \cdots, k\}, \tag{7.37}$$

则有

$$\sigma(C) = \sigma(V^{-1}CV) = \max[u, \sigma(\Delta_1)]. \tag{7.38}$$

这样，由严格根条件，$u < 1$，推得当 $n \to \infty$ 时 $\eta_n \to 0$ 的充要条件为

$$\sigma(\Delta_1) < 1. \tag{7.39}$$

在根 $\zeta_p, p=2, 3, \cdots, k$，有重根的情形，也可以得到同样的结果，只要将块 Vandermonde 矩阵 V 换成广义合块 Vandermonde 矩阵，即换成按下面的步骤得到的矩阵：若 $\zeta_{p^*-1} \neq \zeta_{p^*}$，而 $\zeta_{p^*} =$

ζ_{p^*+1}，则先用 $\zeta_{p^*+\varepsilon}$ 代替矩阵的第 p^*+1 列中的 ζ_{p^*+1}，然后再从第 p^*+1 列减去第 p^* 列并除以 ε，取极限，就得到合块 Vandermonde 矩阵

$$\begin{bmatrix} I & \cdots & \zeta_{p^*}^{k-1}I & (k-1)\zeta_{p^*}^{k-2}I & \cdots \\ I & \cdots & \zeta_{p^*}^{k-2}I & (k-2)\zeta_{p^*}^{k-3}I & \cdots \\ \vdots & & \vdots & \vdots & \\ I & \cdots & \zeta_{p^*}I & I & \cdots \\ I & \cdots & I & 0 & \cdots \end{bmatrix}$$

若有 $\zeta_{p^*}=\zeta_{p^*+1}=\cdots=\zeta_{p^*+v}$，则可重复上述过程得到相应的合块 Vandermonde 矩阵．

由 (7.36)，(7.27) 和引理 7.4 得到

$$\begin{aligned}\Phi_k\Delta_1 &\equiv \left[\sum_{s=0}^{S}\gamma_k^{(s)}(-hA)^s\right]\Delta_1=\sum_{s=0}^{S}\gamma_k^{(s)}(-hA)^s \\ &\quad -\mu_1\sum_{j=0}^{k}\sum_{s=1}^{S}\gamma_j^{(s)}(-hA)^s \\ &=\gamma_k^{(0)}+\sum_{s=1}^{S}\left[\gamma_k^{(s)}+\sum_{l=1}^{s+1}\frac{(-1)^l}{l!}\gamma_k^{(s-l)}\right](-hA)^s \\ &=\sum_{s=0}^{S}\sum_{l=0}^{s}\frac{(-1)^l}{l!}\gamma_k^{(s-l)}(-hA)^s.\end{aligned}$$

因此 $\Delta_1=R_s(hA)$．定理证毕．

如果 $\gamma_k^{(s)}$, $s=0,1,\cdots,S$ 是任意选定的，一般有 $R_s(hA)=e^{hA}+O(h^{S+1})$．但是，对于阶 $p(\geqslant S)$ 的稳定化方法 (7.2)，选取的系数 $\gamma_k^{(s)}$ 满足 (7.21) 的 p 个关系式，这保证有 $R_s(hA)=e^{hA}+O(h^{p+1})$．另外还有 $2S-p$ 个参数可以选成使得 (7.26) 成立．这样的选取是可以做到的．因为在极端的情形，我们可以选取 $\gamma_k^{(s)}$, $s=0,1,\cdots,S$，使得 $R_s(hA)$ 是 e^{hA} 的 (S,S) Padé 近似，即有 $R_s(hA)=e^{hA}+O(h^{2S+1})$，这时 (7.26) 确实是成立的．

在 $S=1,2$ 的情形，按照第五章 §1 例 1.1 中的证明，由定理

7.3 可得到下面的推论.

推论 7.1 对于 $S=1$,阶 $p\geqslant 1$ 的稳定化方法 (7.2) 为 $\bar{A}$ 稳定的充要条件是 $\gamma_k^{(1)}\geqslant\frac{1}{2}$.

推论 7.2 对于 $S=2$,阶 $p\geqslant 2$ 的稳定化方法 (7.2) 为 $\bar{A}$ 稳定的充要条件是 $\gamma_k^{(1)}\geqslant\frac{1}{2}$ 和 $1-2\gamma_k^{(1)}+4\gamma_k^{(2)}\geqslant 0$.

定理 7.3 说明稳定化方法 (7.2) 的 $\bar{A}$ 稳定性只依赖于真解的有理近似的稳定性,寄生根 ζ_p, $p=2,\cdots,k$ 不影响稳定性.

由上面的讨论可以看出, $\bar{A}$ 稳定性是针对试验问题 (7.24) 的数值解的性质. 当方法 (7.2) 应用到一般的初值问题 (7.1) 时,其右函数的 Jacobi 矩阵将随 t 而变化. 我们只能选取 Q_n 为在当前的积分区域中 $-\partial f/\partial y$ 的近似. 另外,为了减少计算量,也不需要每一步重新计算 Q_n. 为了说明对于一般情形,在大的 t 的区域中应用固定步长 h 时,希望方法具有什么样的渐近稳定性质,我们考虑变系数的试验问题

$$y'=A(t)y,\quad y(t_0)=y_0. \tag{7.40}$$

当 $t\to\infty$ 时, (7.40) 的解趋向于零的一个充分条件是对所有 t 有

$$\mu[A(t)]<0, \tag{7.41}$$

其中 $\mu(A)=\lim\limits_{h\to 0^+}(\|I+hA\|-1)/h$ 是矩阵 A 的对数模.

现在还没有得到有关的判别准则,来保证在定理 7.3 的意义下方法 (7.2) 应用到方程 (7.40) 的稳定性. 定理 7.4 给出了方法 (7.2) 应用到方程 (7.40) 的数值解的一个界.

采用定理 7.3 的记号,另外定义

$$B_{S,i}(hQ)=\left[\sum_{s=0}^{S}\gamma_k^{(s)}h^sQ^s\right]^{-1}\left[\sum_{s=0}^{S-1}b_i^{(s)}h^sQ^s\right],$$

$$i=0,1,\cdots,k-1 \tag{7.42}$$

和 $Y_0=\max\|y_i\|_\infty$, $i=0,1,\cdots,k-1$,其中 $\|\cdot\|_\infty$ 表示最大模.

定理 7.4 将形式为 (7.2) 的稳定化线性隐式方法应用到方

程(7.40). 假定方法的特征方程的根是不同的. 令

(i) K, W, L_i, $i=0, 1, 2, \cdots, n$ 是正常数($K \geqslant 1$), 使有

$$\prod_{i=1}^{M} \|R_s(-hQ_{j_i})\|_\infty \leqslant KW^M, \ 0 \leqslant j_1 < j_2 < \cdots < j_M \leqslant n,$$

$$\sum_{j=0}^{k-1} \|B_{s,j}(hQ_i)\|_\infty \cdot \|Q_i + A(t_{i+j})\|_\infty = L_i, \ i=0, 1, \cdots, n;$$

(ii) $\tilde{u}$, u_r, ε_r, $r=2, 3, \cdots, k$ 是正常数, 使有

$$u_r = \frac{1}{\varepsilon_r}\left|\frac{\mu_r}{\mu_1}\right|(KW+1) \geqslant 1, \tilde{u} = \max_r u_r;$$

(iii) α, K^*, K^{**} 是常数, 使有

$$e^{-\alpha} = \max(W, |\zeta_r| + \varepsilon_r, r=2, 3, \cdots, k),$$

$$K^* = \|V\|_\infty \cdot \|V^{-1}\|_\infty \cdot K \cdot \tilde{u}, K^{**} = e^\alpha K^*,$$

于是数值解 y_{n+k} 有界, 即

$$\|y_{n+k}\|_\infty \leqslant K^* Y_0 \exp\left[-\alpha(n+1) + hK^{**}\sum_{i=0}^{n} L_i\right]. \tag{7.43}$$

由(7.43)可以看到, 如果 $\sum_{i=0}^{\infty} L_i < \infty$, 或者 $hK^{**}\max_i L_i < \alpha$, 则当 $n \to \infty$ 时, 有 $y_n \to 0$. 在证明定理之前, 我们需要下面的结果.

引理 7.5 令 $\{y_n\}$ 满足差分方程.

$$y_{n+1} = (C_n + hD_n)y_n,$$

其中 C_n 和 D_n 均是矩阵, 满足

(i) $\left\|\sum_{i=1}^{M} C_{j_i}\right\| \leqslant Ke^{-\alpha M}, \ 0 \leqslant j_1 < j_2 < \cdots < j_M \leqslant n, \ K \geqslant 1,$

(ii) $\|D_i\| = L_i, \ i=0, 1, \cdots, n,$

于是如果 $K^* = Ke^\alpha$, 则有

$$\|y_{n+1}\| \leqslant K\|y_0\|\exp\left[-\alpha(n+1) + hK^*\sum_{i=0}^{n} L_i\right]. \tag{7.44}$$

这个结果是 Strang 引理的推广. 将 Strang [107] 中给出的证明稍作修改可以证明这个引理.

定理 7.4 的证明　将形式为(7.2)的线性隐式方法应用到方程(7.40),得到差分方程组

$$\Phi_k(n)y_{n+k}=\sum_{j=0}^{k-1}[-\Phi_j(n)+hE_j(n)]y_{n+j},\tag{7.45}$$

其中

$$\Phi_j(n)=\sum_{s=0}^{S}\gamma_j^{(s)}h^sQ_n^s,\ j=0,1,\cdots,k,$$

$$E_j(n)=\sum_{s=0}^{S-1}b_j^{(s)}h^sQ_n^s(Q_n+A(t_{n+j})),$$

$$j=0,1,\cdots,k-1.$$

将 (7.45) 写成单步的形式,给出

$$\eta_{n+1}=(C_n+hD_n)\eta_n,$$

其中

$$C_n=\begin{bmatrix}-\Phi_k^{-1}(n)\Phi_{k-1}(n) & -\Phi_k^{-1}(n)\Phi_{k-2}(n) & \cdots & -\Phi_k^{-1}(n)\Phi_0(n)\\ I & 0 & \cdots & 0\\ 0 & I & \cdots & 0\\ \vdots & \ddots & \cdots & 0\\ 0 & & I & 0\end{bmatrix}$$

和

$$D_n=\begin{bmatrix}\Phi_k^{-1}(n)E_{k-1}(n) & \Phi_k^{-1}(n)E_{k-2}(n) & \cdots & \Phi_k^{-1}(n)E_0(n)\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix}.$$

定义

$$\widetilde{V}=V\cdot\text{diag}[I,u_2I,\cdots,u_kI],$$

其中 V 是 Vandermonde 矩阵 (7.31), 而 u_r, $r=2,\cdots,k$ 由本定理给出。

记

$$\bigcap_{i=1}^{M} C_{j_i} = \tilde{V}\left(\bigcap_{i=1}^{M} \tilde{V}^{-1} C_{j_i} \tilde{V}\right) \tilde{V}^{-1},$$

在方法是稳定化的假定下，与在定理 7.3 的证明中一样，我们找到

$$\tilde{V}^{-1} C_{j_i} \tilde{V} = \begin{bmatrix} \Delta_1(j_i) & 0 & \cdots & 0 \\ \frac{1}{u_2}\Delta_2(j_i) & \zeta_2 I & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ \frac{1}{u_k}\Delta_k(j_i) & 0 & \cdots & \zeta_k I \end{bmatrix}, \tag{7.46}$$

其中

$$\Delta_1(j_i) = R_s(-hQ_{j_i}),$$

$$\Delta_r(j_i) = \frac{\mu_r}{\mu_1}(\Delta_1(j_i) - I), \quad r = 2, 3, \cdots, k.$$

因此，由定理的假定 (i) 和 (ii)

$$\|\tilde{V}^{-1} C_{j_i} \tilde{V}\|_\infty \leqslant \max\begin{cases} \|R_s(-hQ_{j_i})\|_\infty \leqslant KW \\ \frac{1}{u_r}\left|\frac{\mu_r}{\mu_1}\right|(KW+1) + |\zeta_r| \leqslant \varepsilon_r + |\zeta_r|, \\ \qquad r = 2, 3, \cdots, k. \end{cases} \tag{7.47}$$

所以

$$\prod_{i=1}^{M} \|\tilde{V}^{-1} C_{j_i} \tilde{V}\|_\infty \leqslant \max\begin{cases} KW^M, \\ K(\varepsilon_r + |\zeta_r|)^M, \ r = 2, 3, \cdots, k. \end{cases} \tag{7.48}$$

对于$K = 1$的情形，不等式 (7.48) 立即由 (7.47) 得到. 如果用 $K = K^* > 1$ 代替 $K = 1$，则由定理的假定 (i) 和矩阵 (7.46)的形式，(7.48) 中的模最多增加一个因子 K^*. 因此不等式 (7.48)对于$K > 1$ 也成立. 最后，由于 $u_r \geqslant 1$, $r = 2, 3, \cdots, k$,

$$\|\tilde{V}\|_\infty \leqslant \|V\|_\infty \cdot \|\mathrm{diag}[I, u_2 I, \cdots, u_k I]\|_\infty = \|V\|_\infty \tilde{u},$$

$$\|\tilde{V}^{-1}\|_\infty \leqslant \left\|\mathrm{diag}\left[I, \frac{1}{u_2} I, \cdots, \frac{1}{u_k} I\right]\right\|_\infty \cdot \|V^{-1}\|_\infty = \|V^{-1}\|_\infty.$$

因此，由定理的假定 (iii)，

$$\left\|\prod_{i=1}^{M} C_{i_i}\right\|_\infty \leqslant \|V\|_\infty \tilde{u} K e^{-\alpha M} \|V^{-1}\|_\infty,$$

还有

$$\|D_i\|_\infty \leqslant \sum_{j=0}^{k-1} \|B_{s,j}(hQ_i)\|_\infty \cdot \|Q_i + A(t_{i+j})\|_\infty,$$

由引理 7.5，定理得证。

对于方法的特征多项式具有重根的情形，可以得到类似的定理，但证明中用到的 Vandermonde 矩阵必须换成广义合块 Vandermonde 矩阵。

在完全隐式方法的情形，稍复杂一些。可以将引理 7.5 作下面的推广，再进行处理。

令 $\{y_n\}$ 满足差分方程

$$(I + hE_n)y_{n+1} = (C_n + hD_n)y_n,$$

其中 $(I + hE_n)$ 非奇异，于是

$$y_{n+1} = [C_n + h(I + hE_n)^{-1}(-E_nC_n + D_n)]y_n$$

若将引理 7.5 的假定 (ii) 换成

$$\|(I + hE_i)^{-1}(-E_iC_i + D_i)\| = L_i, \ i = 0, 1, 2, \cdots, n,$$

则引理 7.5 的结论仍成立。

因此，对于完全隐式方法，只要将定理 7.4 的假定 (ii) 中 L_i 的表达式作一些适当的改变，其余的假定可以保持不变。

§4 $\bar{A}$ 稳定方法的例子

这一节给出一些形式为 (7.2) 的 $\bar{A}$ 稳定方法类。除方法 7.1 和 7.4 外，所有的方法均是线性隐式的。方法中出现二族不同的参数类 α, β 和 a, b, c. 下面给出了为保证 $\bar{A}$ 稳定性参数必须满足的条件。除这些条件外，参数是自由的。对于每一类方法，还给出 (7.25) 的 $R_t(hA)$ 和局部截断误差 $(T.E)$。

$$T.E. = \tilde{L}[y(t); h],$$

方法 7.1　$S=1$，$k=1$，$p=1, 2$，

$$a_1^{(0)}=1,\quad a_0^{(0)}=-1,$$
$$b_1^{(0)}=b,\quad b_0^{(0)}=1-b,$$
$$a_1^{(1)}=a,\quad a_0^{(1)}=-a,$$

如果 $a+b\geqslant\frac{1}{2}$，方法是 $\bar{A}$ 稳定的．

$$R_1(hA)=[I-(a+b)hA]^{-1}[I-(a+b-1)hA].$$

$$T.E.=h^2\left\{\left(\frac{1}{2}-b\right)y^2(t)+aQy^{(1)}(t)\right\}+O(h^3)$$

如果 $b=\frac{1}{2}$，$a=0$，则有

$$T.E.=-\frac{1}{12}h^3y^{(3)}(t)+O(h^4).$$

若 $b=0$，给出线性隐式方法．$b=\frac{1}{2}$，$a=0$ 给出梯形法．

方法 7.2　$S=1$，$k=2$，$p=2$

$$a_2^{(0)}=1,\quad a_1^{(0)}=-1-\alpha,\quad a_0^{(0)}=\alpha,$$
$$b_2^{(0)}=0,\quad b_1^{(0)}=\frac{1}{2}(3-\alpha),\quad b_0^{(0)}=-\frac{1}{2}(1+\alpha),$$
$$a_2^{(1)}=\frac{1}{2},\quad a_1^{(1)}=-1,\quad a_0^{(1)}=\frac{1}{2}.$$

这类方法都是线性隐式的．如果 $-1<\alpha<1$，方法是 $\bar{A}$ 稳定的．

$$R_1(hA)=\left[I-\frac{1}{2}hA\right]^{-1}\left[I+\frac{1}{2}hA\right],$$

$$T.E.=h^3\left\{\frac{1}{12}(5+\alpha)y^{(3)}(t)+\frac{1}{2}Qy^{(2)}(t)\right\}+O(h^4).$$

方法 7.3　$S=2$，$k=2$，$p=2$

$$a_2^{(0)}=1,\quad a_1^{(0)}=-1-\alpha,\quad a_0^{(0)}=\alpha,$$
$$b_2^{(0)}=0,\quad b_1^{(0)}=\frac{1}{2}(3-\alpha),\quad b_0^{(0)}=-\frac{1}{2}(1+\alpha),$$
$$a_2^{(1)}=a,\quad a_1^{(1)}=-\frac{1}{2}(1-\alpha)-a(1+\alpha),$$

$$a_0^{(1)} = \frac{1}{2}(1-\alpha) + a\alpha, \quad b_2^{(1)} = 0,$$

$$b_1^{(1)} = -\frac{1}{2} + a - b(1+\alpha) - c,$$

$$b_0^{(1)} = \frac{1}{2}\alpha - a\alpha + b(1+\alpha) + c,$$

$$a_2^{(2)} = b, \quad a_1^{(2)} = c, \quad a_0^{(2)} = -b - c.$$

这类中的方法都是线性隐式的. 如果 $-1<\alpha<1$, $4b \geqslant 2a-1 \geqslant 0$,方法是 $\bar{A}$ 稳定的.

$$R_2(hA) = [I - ahA + bh^2A^2]^{-1} \cdot \left[I - (a-1)hA + \left(\frac{1}{2} - a + b\right)h^2A^2\right]$$

$$T.E. = h^3\left\{\frac{1}{12}(5+\alpha)y^{(3)}(t) + \left[\frac{1}{4}(1+\alpha) + \frac{1}{2}a(1-\alpha) + b(1+\alpha) + c\right]Qy^{(2)}(t) + (2b+c)Q^2y^{(1)}(t)\right\} + O(h^4).$$

这类方法达到与方法 7.2 同样的阶,但是在方法的形式中含有更多的自由参数供使用者选取. 当 $a=\frac{1}{2}$, $b=c=0$ 时,又得到方法 7.2.

方法 7.4 $S=2$, $k=2$, $p=3, 4$

$$a_2^{(0)} = 1, \quad a_1^{(0)} = -1-\alpha, \quad a_0^{(0)} = \alpha,$$

$$b_2^{(0)} = \frac{1}{12}(5+\alpha), \quad b_1^{(0)} = \frac{2}{3}(1-\alpha),$$

$$b_0^{(0)} = -\frac{1}{12}(1+5\alpha), \quad a_2^{(1)} = 2a - \frac{1}{12}(1+\alpha),$$

$$a_1^{(1)} = \frac{1}{3}\alpha - 2a(1+\alpha), \quad a_0^{(1)} = \frac{1}{12}(1-3\alpha) + 2a\alpha,$$

$$b_2^{(1)} = a - b, \quad b_1^{(1)} = -\frac{1}{6} + a(1-\alpha) + 2b,$$

$$b_0^{(1)} = \frac{1}{6}\alpha - a\alpha - b, \quad a_2^{(2)} = b, \quad a_1^{(2)} = -2b,$$

$$a_0^{(2)} = b.$$

这是一个完全隐式方法类. 满足 $-1 < \alpha < 1, a \geqslant \frac{1}{2}$ 的方法是 $\bar{A}$ 稳定的.

$$R_2(hA) = \left[I - \left(\frac{1}{3} + 2a\right)hA + ah^2A^2\right]^{-1}$$

$$\cdot \left[I + \left(\frac{2}{3} - 2a\right)hA + \left(\frac{1}{6} - a\right)h^2A^2\right],$$

$$T.E. = h^4\left\{-\frac{1}{24}(1+\alpha)y^{(4)}(t) + \left[-\frac{1}{36}(1+2\alpha)\right.\right.$$

$$\left.\left.-\frac{1}{6}a(1-\alpha) + b\right]Qy^{(3)}(t) + bQ^2y^{(2)}(t)\right\} + O(h^5).$$

如果 $\alpha = -1, a = \frac{1}{12}, b = 0$,则

$$T.E. = h^5\left[-\frac{1}{90}y^{(5)}(t) - \frac{1}{72}Qy^{(4)}(t)\right] + O(h^6).$$

由上面 $T.E.$ 的公式,为得到 4 阶的方法,必须取 $\alpha = -1$,得到的方法是 Simpson 法则的稳定化格式. 由于不满足严格的根条件,它不是 $\bar{A}$ 稳定的. 在这种极限情形,当 $n \to \infty$ 时,寄生解仍是有界的,但不趋向于零.

方法 7.5 $S = 2, k = 3, p = 3,$

$$a_3^{(0)} = 1, \qquad a_2^{(0)} = -1 - \alpha,$$

$$a_1^{(0)} = \alpha + \beta, \qquad a_0^{(0)} = -\beta,$$

$$b_3^{(0)} = 0, \qquad b_2^{(0)} = \frac{1}{12}(23 - 5\alpha - \beta),$$

$$b_1^{(0)} = -\frac{2}{3}(2 + \alpha - \beta),$$

$$b_0^{(0)} = \frac{1}{12}(5 + \alpha + 5\beta),$$

$$a_3^{(1)} = \frac{1}{3} + 2a,$$

$$a_2^{(1)} = \frac{1}{12}[-15 + \alpha + \beta - 24a(1 + \alpha)],$$

$$a_1^{(1)} = \frac{1}{3}[4 - \beta + 6a(\alpha + \beta)],$$

$$a_0^{(1)} = \frac{1}{12}[-5 - \alpha + 3\beta - 24a\beta],$$

$$b_3^{(1)} = 0, \qquad b_2^{(1)} = -\frac{1}{6} + a(1 - \alpha) - b,$$

$$b_1^{(1)} = \frac{1}{6}\alpha + a(3 - \alpha + \beta) + 2b,$$

$$b_0^{(1)} = -\frac{1}{6}\beta - a(2 - \beta) - b,$$

$$a_3^{(2)} = a, \qquad a_2^{(2)} = b,$$

$$a_1^{(2)} = -3a - 2b, \qquad a_0^{(2)} = 2a + b.$$

这是一个线性隐式的方法类． 若将 α, β 记成 $\alpha = \zeta_2 + \zeta_3$, $\beta = \zeta_2\zeta_3$，则如果选取 α, β 使有 $|\zeta_i| < 1$, $i = 2, 3$ 和如果 $a \geqslant \frac{1}{12}$，方法是 $\bar{A}$ 稳定的.

$$R_2(hA) = \left[I - \left(\frac{1}{3} + 2a\right)hA + ah^2A^2\right]^{-1} \cdot \left[I + \left(\frac{2}{3} - 2a\right)hA + \left(\frac{1}{6} - a\right)h^2A^2\right],$$

$$T.E. = h^4\left\{\frac{1}{24}(9 + \alpha + \beta)y^{(4)}(t) + \left[\frac{1}{36}(14 + \alpha + 2\beta) + \frac{1}{6}a(17 + \alpha - \beta) + b\right]Qy^{(3)}(t) + (3a + b)Q^2y^{(2)}(t)\right\} + O(h^5)$$

由 $T.E.$ 的形式及 $\bar{A}$ 稳定性的 α, β 的取值范围可知不存在 α 和 β 的值使方法的阶超过了，并且保持 $\bar{A}$ 稳定性．为了达到阶 4，必须考虑 $S = 2$ 和 $k = 3$（完全隐式）或 $k = 4$（线性隐式）的情形．

下面是 $\bar{A}$ 稳定的线性隐式方法的一个特殊例子，其中选取的自由参数使给出了方便的系数.

方法 7.6 $S=2$, $k=4$, $p=4$ 的特殊的 $\bar{A}$ 稳定线性隐式方法.

$$a_4^{(0)}=1,\ a_3^{(0)}=-1,\ a_2^{(0)}=0,\ a_1^{(0)}=0,\ a_0^{(0)}=0,$$

$$b_4^{(0)}=0,\ b_3^{(0)}=\frac{55}{24},\ b_2^{(0)}=-\frac{59}{24},\ b_1^{(0)}=\frac{37}{24},\ b_0^{(0)}=-\frac{9}{24},$$

$$a_4^{(1)}=\frac{1}{2},\ a_3^{(1)}=-\frac{43}{24},\ a_2^{(1)}=\frac{59}{24},\ a_1^{(1)}=-\frac{37}{24},\ a_0^{(1)}=\frac{9}{24},$$

$$b_4^{(1)}=0,\ b_3^{(1)}=\frac{1}{6},\ b_2^{(1)}=-\frac{1}{4},\ b_1^{(1)}=\frac{1}{12},\ b_0^{(1)}=0,$$

$$a_4^{(2)}=\frac{1}{12},\ a_3^{(2)}=-\frac{1}{4},\ a_2^{(2)}=\frac{1}{4},\ a_1^{(2)}=-\frac{1}{12},\ a_0^{(2)}=0,$$

$$R_2(hA)=\left[I-\frac{1}{2}hA+\frac{1}{12}h^2A^2\right]^{-1}\cdot\left[I+\frac{1}{2}hA+\frac{1}{12}h^2A^2\right]$$

$$T.E.=h^5\left\{\frac{251}{720}y^{(5)}(t)+\frac{31}{72}Qy^{(4)}(t)+\frac{1}{12}Q^2y^{(3)}(t)\right\}+O(h^6).$$

本 章 附 注

本章的材料主要取自 Lambert 和 Sigurdsson 的[70].

第八章 边界层方法

§1 奇异摄动问题的解的渐近展开式

奇异摄动的微分方程组是一类很重要的刚性方程. 通常用来求解奇异摄动组的解析方法称作边界层方法. 可以利用这些方法来构造求解刚性组的数值方法. 按照这种思想推导的方法称为边界层型数值方法.

这一节我们将简单地介绍 Тихонов 的极限过渡定理和 Васильева 的构造奇异摄动组的解的渐近展开式的算法. 它们是求解许多问题的基础.

考虑常微分方程组

$$\begin{aligned} \mu\frac{dz}{dt} &= F(z, y, t), \\ \frac{dy}{dt} &= f(z, y, t), \end{aligned} \tag{8.1}$$

其中 z 和 F 是 M 维向量函数, y 和 f 是 m 维向量函数, $\mu > 0$ 是小参数.

给定初始条件(为简单起见, 令 $t_0 = 0$)

$$z(0, \mu) = z_0, \quad y(0, \mu) = y_0. \tag{8.2}$$

这里认为 z_0 和 y_0 不依赖于参数 μ. 设问题 (8.1) 和 (8.2) 在区间 $0 \leqslant t \leqslant T$ 上的解为 $z(t, \mu)$, $y(t, \mu)$.

如果令 (8.1) 中的 μ 为零, 得到

$$0 = F(\bar{z}, \bar{y}, t), \tag{8.3a}$$

$$\frac{d\bar{y}}{dt} = f(\bar{z}, \bar{y}, t), \tag{8.3b}$$

称它为 (8.1) 的退化组, 而 (8.1) 称作完全组或称作奇异摄动组.

y 称作 (8.1) 的正则部分，而 z 称作奇异部分. 由 (8.3) 可以看出 (8.1) 的前M个方程退化成 (8.3) 中的代数方程. 因此 (8.3) 的阶比 (8.1) 的小，(8.3) 所需要的初始条件的个数也比 (8.1) 的少. 对于 (8.3) 的初始条件的最自然的取法为

$$\bar{y}(0) = y_0, \tag{8.4}$$

而将对 z 的初始条件舍弃掉.

为了求解 (8.3),对于 $\bar{y}$, t 由 (8.3a) 解出 $\bar{z} = \varphi(\bar{y}, t)$ (由于 F 一般是非线性的,这样求得的 $\bar{z}$ 可能不是唯一的,可以利用连续性或其它的性质从中选取所需要的 $\bar{z}$ 的值.)，将其代入 (8.3b)，得到初值问题

$$\frac{d\bar{y}}{dt} = f(\varphi(\bar{y}, t), \bar{y}, t),\ \bar{y}(0) = y_0, \tag{8.5}$$

由它可以求得 $\bar{y}(t)$ 以及 $\bar{z}(t) = \varphi(\bar{y}(t), t)$. 称它们为退化解.

通常按上述方式求得的 $\bar{z}(t)$ 将不满足 (8.2) 中对 z 的初始条件，即 $\bar{z}(0) \neq z_0$. 因此，至少在 $t = 0$ 的某个邻域中退化解 $\bar{z}(t)$ 将不能任意接近于完全组 (8.1) 的解 $z(t, \mu)$. 但是要问在这个邻域的外面 $\bar{z}(t)$ 是否能任意接近解 $z(t, \mu)$ 和在 $0 < t \leqslant T$ 上退化解 $\bar{y}(t)$ 是否能任意接近 $y(t, \mu)$? 按照 (8.1)、(8.2) 所具有的条件和解 $\bar{z} = \varphi(\bar{y}, t)$ 的选取方式，这个问题的答案可以是肯定的，也可以是否定的. 下面的定理 8.1 将给出一个肯定的结果.

为了给出定理 8.1,我们叙述下面的条件:

I　在变量 (z, y, t) 空间的某个开区域G中，函数 $F(z, y, t)$ 和 $f(z, y, t)$ 连续并满足对 z 和 y 的 Lipschitz 条件.

II　在变量 (y, t) 空间的某个有界闭域 $\bar{D}$ 中对于 (z, t) 方程 $F(z, y, t) = 0$ 有解(根) $z = \varphi(y, t)$，且具有性质

1. $\varphi(y, t)$ 是在 $\bar{D}$ 中的连续函数.

2. 当 $(y, t) \in \bar{D}$ 时,点 $(\varphi(y, t), y, t) \in G$.

3. 在 $\bar{D}$ 中根 $z = \varphi(y, t)$ 是孤立的,即存在 $\eta > 0$ 使当 $0 < \|z - \varphi(y, t)\| < \eta$，$(y, t) \in \bar{D}$ 时，$F(z, y, t) \neq 0$.

III (8.5) 在区间 $0 \leqslant t \leqslant T$ 上有唯一解 $\bar{y}(t)$，并且当 $t \in [0, T]$ 时，点 $(\bar{y}(t), t) \in D$，这里 D 是区域 $\bar{D}$ 的内点的集合。另外，假定在 $\bar{D}$ 中函数 $f(\varphi(y, t), y, t)$ 满足对 y 的 Lipschitz 条件.

现在引进辅助微分方程组

$$\frac{d\tilde{z}}{d\tau} = F(\tilde{z}, y, t) \quad (\tau \geqslant 0), \tag{8.6}$$

其中 y 和 t 看成是参数. 由 II，当 $(y, t) \in \bar{D}$ 时，$\tilde{z} = \varphi(y, t)$ 是 (8.6) 的孤立静止点. 构造条件

IV (8.6) 的静止点 $\tilde{z} = \varphi(y, t)$ 对于 $(y, t) \in D$ 是按 Ляпунов 一致渐近稳定的.

这表示对于任意的 $\varepsilon > 0$，存在 $\delta(\varepsilon)$，使当 $\|\tilde{z}(0) - \varphi(y, t)\| < \delta(\varepsilon)$ 时，有 $\|\tilde{z}(\tau) - \varphi(y, t)\| < \varepsilon$. 当 $\tau \geqslant 0$ 和 $\tau \to \infty$ 时，$\tilde{z}(\tau) \to \varphi(y, t)$.

在条件 IV 满足的情形，根 $\tilde{z} = \varphi(y, t)$ 称作在区域 $\bar{D}$ 中稳定.

现在考虑 $y = y_0$，$t = 0$ 时的辅助 (8.6).

$$\frac{d\tilde{z}}{d\tau} = F(\tilde{z}, y_0, 0) \quad (\tau \geqslant 0), \tag{8.7}$$

并取其初始条件为

$$\tilde{z}(0) = z_0. \tag{8.8}$$

因为一般来说初始值 z_0 不接近于静止点 $\varphi(y_0, 0)$，因而当 $\tau \to \infty$ 时，问题 (8.7)、(8.8) 的解 $\tilde{z}(t)$ 可能不趋向于 $\varphi(y_0, 0)$. 但我们提出条件

V 问题 (8.7)、(8.8) 的解 $\tilde{z}(\tau)$ 满足条件

1. $\tilde{z}(\tau) \to \varphi(y_0, 0)$，当 $\tau \to \infty$ 时，
2. 点 $(\tilde{z}(\tau), y_0, 0) \in G$，当 $\tau \geqslant 0$ 时.

在这种情形，将称初始点 z_0 属于静止点 $\tilde{z} = \varphi(y_0, 0)$ 的影响区域. 对于渐近稳定的静止点 $\tilde{z} = \varphi(y_0, 0)$，至少对于所有充分接近它的点均属于它的影响区域.

Тихонов 给出下面的定理:

定理 8.1 (Тихонов 定理) 设条件 I—V 成立，则可找到常数 $\mu_0 > 0$，使当 $0 < \mu \leqslant \mu_0$ 时，在区间 $0 \leqslant t \leqslant T$ 上问题 (8.1)、(8.2) 的解 $z(t, \mu)$，$y(t, \mu)$ 存在，唯一并且满足极限等式

$$\lim_{\mu \to 0} y(t, \mu) = \bar{y}(t), \quad 当\ 0 \leqslant t \leqslant T\ 时, \tag{8.9}$$

$$\lim_{\mu \to 0} z(t, \mu) = \bar{z}(t) = \varphi(\bar{y}(t), t), \quad 当\ 0 < t \leqslant T\ 时. \tag{8.10}$$

这个定理的详细证明见 Васильева 和 Бутузов [114]. 从证明中还可以看出 $y(t, \mu)$ 的极限过程 (8.9) 关于 $t \in [0, T]$ 是一致的,而 $z(t, \mu)$ 的极限过程 (8.10) 对于 $t \in [0, T]$ 不是一致的,但对于 $t \in [\bar{t}_0, T]$ 是一致的,其中 $\bar{t}_0 > 0$ 可以任意小,但当 $\mu \to 0$ 时是固定的.

定理 8.1 说明退化组 (8.3)、(8.4) 的解 $\bar{z}(t)$，$\bar{y}(t)$ 与完全组 (8.1)、(8.2) 的解 $z(t, \mu)$，$y(t, \mu)$ 之间的关系. 当 $\mu > 0$ 很小时可以用 $\bar{y}(t)$ 作为 $y(t, \mu)$ 的近似,并且在区间 $0 \leqslant t \leqslant T$ 上具有一致的精度. 对于 $\bar{z}(t)$ 情形就不同了. 在 $t = 0$ 处，$\bar{z}(t)$ 与 $z(t, \mu)$ 之间的差为 $z_0 - \bar{z}(0)$，一般说来是不同的. 于是在 $t = 0$ 的某个邻域中，$\bar{z}(t)$ 不能作为 $z(t, \mu)$ 的近似. 但是在这个 $t = 0$ 的小的邻域中 $z(t,\mu)$迅速地从 z_0 变化到接近于 $\bar{z}(t)$ 的值. 这种现象就是边界层现象,而这个小邻域为边界层区域(简称边界层). 而在区间 $0 < \bar{t}_0 \leqslant t \leqslant T$ 上 $\bar{z}(t)$ 可以作为 $\mu \to 0$ 的 $z(t,\mu)$ 的一致渐近的近似. 这里 $\bar{t}_0$ 是固定的可以取得任意小的值.

但是定理 8.1 并未给出这些渐近近似的精度,只给出一个定性的结果. 自然想得到在区间 $0 \leqslant t \leqslant T$ 可以有以任意精度的 $y(t, \mu)$ 和 $z(t, \mu)$ 的一致的近似. 为此在下面给出一个算法,使对任意整数 n 给出具有精度 $O(\mu^{n+1})$ 的 $z(t, \mu)$，$y(t, \mu)$ 对于 $t \in [0, T]$为一致的渐近近似. 因此,需要将定理 8.1 中的条件加强. 将条件 I 改成:

I′ 设函数 $F(z, y, t)$ 和 $f(z, y, t)$ 在区域G中对所有变量具有所需要的任意阶连续偏导数.

定理 8.1 中的条件 II、III、V 不变仍记为 II、III、V.

定理 8.1 的条件 IV 是要求辅助组

$$\frac{d\tilde{z}}{d\tau}=F(\tilde{z}, y, t) \quad (y, t \text{ 为参数})$$

的孤立静止点 $\tilde{z}=\varphi(y, t)$ 对于 $\overline{D}$ 是一致渐近稳定的. 为了建立下面的近似,我们需要更具体的要求. 用 $\lambda_i(y, t)(i=1, 2, \cdots, M)$ 表示矩阵 $F_z(\varphi(y, t), y, t)=\left(\frac{\partial F^i}{\partial z^j}\right)_{z=\varphi(y,t)}$ 的特征值, 而用 $\bar{\lambda}_i(t)$ 表示矩阵 $\overline{F}_z(t)\equiv F_z(\varphi(\bar{y}(t), t), \bar{y}(t), t)$ 的特征值, 即有 $\bar{\lambda}_i(t)=\lambda_i(\bar{y}(t), t)$, 这里 $\bar{y}(t)$, $\bar{z}(t)=\varphi(\bar{y}(t), t)$ 是退化组 (8.3)、(8.4) 的解. 引进条件 IV′ 为

IV′ 设成立

$$\operatorname{Re}\bar{\lambda}_i(t)<0, \text{ 当 } 0\leqslant t\leqslant T, \ i=1, \cdots, M. \tag{8.11}$$

称 (8.11) 为稳定性条件.

由于 $F_z(\varphi(y, t), y, t)$ 的连续性推得 $\lambda_i(y, t)$ 的连续性, 存在正数 $\eta>0$, 使区域 $\overline{D}_1=\{(y, t): \|y-\bar{y}(t)\|\leqslant\eta, 0\leqslant t\leqslant T\}\subseteq\overline{D}$, 并且对于 $(y, t)\in\overline{D}_1$, $i=1, 2, \cdots, M$ 有

$$\operatorname{Re}\lambda_i(y, t)<-\alpha<0 \tag{8.12}$$

($\alpha>0$ 是某个常数). 由此推出 $\tilde{z}=\varphi(y, t)$ 是辅助组相对于 $\overline{D}_1$ 为一致的渐近稳定静止点(这样定理 8.1 的条件 IV 仍满足).

现在我们给出一种算法,由它可以建立问题 (8.1)、(8.2) 的解的渐近分解式. 设要建立的渐近分解式具有形式

$$x(t, \mu)=\bar{x}(t, \mu)+\Pi x(\tau, \mu) \quad (\tau=t/\mu), \tag{8.13}$$

其中

$$\bar{x}(t, \mu)=\bar{x}_0(t)+\mu\bar{x}_1(t)+\cdots+\mu^k\bar{x}_k(t)+\cdots, \tag{8.14}$$

$$\Pi x(\tau, \mu)=\Pi_0x(\tau)+\mu\Pi_1x(\tau)+\cdots+\mu^k\Pi_kx(\tau)+\cdots. \tag{8.15}$$

这里及后面将用 x 来表示 y 和 z, 即如果对 x 满足某个关系

式．这表示对 y 和 z 将满足同样的表达式．$\Pi_k x(\tau)$ 为边界函数．记号 Π 将用来表示不同的函数：$\Pi_k z$，$\Pi_k y$，$\Pi_k F$，$\Pi_k f$ 等．它们的相似之处是仅在初始点的某个小邻域中存在值，随着 τ 的增大将迅速衰减．对于每个 k，记号 Π_k 可以看成是作用在任意函数 (x, F, f) 上的确定的算子．$\bar{x}, \bar{x}_0, \cdots, \bar{x}_k, \cdots$ 称作外部解或边界层外解．

将 (8.13) 代入 (8.1)，并且为对称起见将 (8.1) 的第二个方程乘上 μ，得到

$$
\begin{aligned}
\mu \frac{d\bar{z}}{dt} + \frac{d\Pi z}{d\tau} &= F(\bar{z} + \Pi z, \bar{y} + \Pi y, t), \\
\mu \frac{d\bar{y}}{dt} + \frac{d\Pi y}{d\tau} &= \mu f(\bar{z} + \Pi z, \bar{y} + \Pi y, t).
\end{aligned} \tag{8.16}
$$

将 (8.16) 的右边恒等地变换成类似于 (8.13) 的形式． 现在对 F 进行这种变换(对 f 是完全一样的)．

$$
\begin{aligned}
F(\bar{z} + \Pi z, \bar{y} + \Pi y, t) &= F(\bar{z}(t, \mu), \bar{y}(t, \mu), t) \\
&\quad + F(\bar{z}(\tau\mu, \mu) + \Pi z(\tau, \mu), \bar{y}(\tau\mu, \mu) \\
&\quad + \Pi y(\tau, \mu), \tau\mu) - F(\bar{z}(\tau\mu, \mu), \bar{y}(\tau\mu, \mu), \tau\mu) \\
&\equiv \bar{F} + \Pi F.
\end{aligned}
$$

将分解式 (8.14)、(8.15) 代入式中的 $\bar{x}$ 和 Πx，将 $\bar{F}$ 和 ΠF 表成 μ 的幂级数的形式

$$
\begin{aligned}
\bar{F} &\equiv F(\bar{z}(t, \mu), \bar{y}(t, \mu), t) \\
&= F(\bar{z}_0(t) + \mu\bar{z}_1(t) + \cdots + \mu^k\bar{z}_k(t) + \cdots, \bar{y}_0(t) \\
&\quad + \mu\bar{y}_1(t) + \cdots + \mu^k\bar{y}_k(t) + \cdots, t) \\
&= F(\bar{z}_0(t), \bar{y}_0(t), t) + \mu[\bar{F}_z(t)\bar{z}_1(t) + \bar{F}_y(t)\bar{y}_1(t)] \\
&\quad + \cdots + \mu^k[\bar{F}_z(t)\bar{z}_k(t) + \bar{F}_y(t)\bar{y}_k(t) + F_k(t)] + \cdots \\
&\equiv \bar{F}_0 + \mu\bar{F}_1 + \cdots + \mu^k\bar{F}_k + \cdots,
\end{aligned} \tag{8.17}
$$

其中矩阵 $\bar{F}_z(t) = \left(\dfrac{\partial F^i}{\partial z^j}\right)$ 和 $\bar{F}_y(t) = \left(\dfrac{\partial F^i}{\partial y^j}\right)$ 的元素均在点 $(\bar{z}_0(t), \bar{y}_0(t), t)$ 上计算，向量 $F_k(t)$ 是由 $\bar{z}_i(t)$，$\bar{y}_i(t)$ $(i = 1, \cdots, k-1)$ 的确定的表达式计算．

$$
\begin{aligned}
\Pi F \equiv & F(\bar{z}(\tau\mu, \mu) + \Pi z(\tau, \mu), \bar{y}(\tau\mu, \mu) \\
& + \Pi y(\tau, \mu), \tau\mu) - F(\bar{z}(\tau\mu, \mu), \bar{y}(\tau\mu, \mu), \tau\mu) \\
= & F(\bar{z}_0(\tau\mu) + \mu\bar{z}_1(\tau\mu) + \cdots + \mu^k\bar{z}_k(\tau\mu) + \cdots \\
& + \Pi_0 z(\tau) + \mu\Pi_1 z(\tau) + \cdots + \mu^k\Pi_k z(\tau) \\
& + \cdots, \bar{y}_0(\tau\mu) + \mu\bar{y}_1(\tau\mu) + \cdots + \mu^k\bar{y}_k(\tau\mu) \\
& + \cdots + \Pi_0 y(\tau) + \mu\Pi_1 y(\tau) + \cdots + \mu^k\Pi_k y(\tau) \\
& + \cdots, \tau\mu) - F(\bar{z}_0(\tau\mu) + \mu\bar{z}_1(\tau\mu) \\
& + \cdots + \mu^k\bar{z}_k(\tau\mu) + \cdots, \bar{y}_0(\tau\mu) + \mu\bar{y}_1(\tau\mu) \\
& + \cdots + \mu^k\bar{y}_k(\tau\mu) + \cdots, \tau\mu) \\
= & [F(\bar{z}_0(0) + \Pi_0 z(\tau), \bar{y}_0(0) + \Pi_0 y(\tau), 0) \\
& - F(\bar{z}_0(0), \bar{y}_0(0), 0] + \mu[F_z(\tau)\Pi_1 z(\tau) \\
& + F_y(\tau)\Pi_1 y(\tau) + G_1(\tau)] + \cdots \\
& + \mu^k[F_z(\tau)\Pi_k z(\tau) + F_y(\tau)\Pi_k y(\tau) \\
& + G_k(\tau)] + \cdots \\
\equiv & \Pi_0 F + \mu\Pi_1 F + \cdots + \mu^k\Pi_k F + \cdots,
\end{aligned} \tag{8.18}
$$

其中矩阵 $F_z(\tau)$ 和 $F_y(\tau)$ 的元素在点$(\bar{z}_0(0)+\Pi_0 z(\tau), \bar{y}_0(0)+\Pi_0 y(\tau), 0)$ 处计算，而向量 $G_k(\tau)$ 由 $\Pi_i x(\tau)$ $(i=0, 1, \cdots, k-1)$ 的表达式确定.

对于函数 f 具有同样的分解式. (8.18) 的最后的恒等式作为算子 $\Pi_k(k=0, 1, 2, \cdots)$ 的定义.

将 $F=\bar{F}+\Pi F$ 和 $f=\bar{f}+\Pi f$ 代入 (8.16),得到

$$
\begin{aligned}
\mu\frac{d\bar{z}}{dt} + \frac{d\Pi z}{d\tau} &= \bar{F} + \Pi F, \\
\mu\frac{d\bar{y}}{dt} + \frac{d\Pi y}{d\tau} &= \mu\bar{f} + \mu\Pi f.
\end{aligned} \tag{8.19}
$$

用分解式 (8.14)、(8.15) 代替 (8.19) 式左边的 $\bar{z}, \bar{y}, \Pi z, \Pi y$，而用分解式 (8.17)、(8.18) 和 $\bar{f}$, Πf 的类似的分解式代替 (8.19) 的右边部分. 然后分别按照依赖于 t 或 τ 让等式两边的 μ 的同次幂的系数相等，于是得到确定分解式 (8.14)、(8.15) 的各个项的方程.

对于零阶近似(即 $\bar{x}_0(t)$, $\Pi_0 x(\tau)$), 有

$$\begin{aligned} 0 &= \bar{F}_0 \equiv F(\bar{z}_0, \bar{y}_0, t), \\ \frac{d\bar{y}_0}{dt} &= \bar{f}_0 \equiv f(\bar{z}_0, \bar{y}_0, t), \end{aligned} \tag{8.20}$$

这恰好与退化组 (8.3) 复合.

$$\begin{aligned} \frac{d\Pi_0 z}{d\tau} &= \Pi_0 F \equiv F(\bar{z}_0(0) + \Pi_0 z, \bar{y}_0(0) + \Pi_0 y, 0) \\ &\quad - F(\bar{z}_0(0), \bar{y}_0(0), 0) \\ &= F(\bar{z}_0(0) + \Pi_0 z, \bar{y}_0(0) + \Pi_0 y, 0), \\ \frac{d\Pi_0 y}{d\tau} &= 0 \end{aligned} \tag{8.21}$$

(由于 (8.20), $F(\bar{z}(0), \bar{y}_0(0), 0) = 0$).

对于分解式 (8.14)、(8.15) 的足标为 1 的项,有方程

$$\begin{aligned} \frac{d\bar{z}_0}{dt} &= \bar{F}_1 \equiv \bar{F}_z(t)\bar{z}_1 + \bar{F}_y(t)\bar{y}_1, \\ \frac{d\bar{y}_1}{dt} &= \bar{f}_1 \equiv \bar{f}_z(t)\bar{z}_1 + \bar{f}_y(t)\bar{y}_1, \end{aligned} \tag{8.22}$$

$$\begin{aligned} \frac{d\Pi_1 z}{d\tau} &= \Pi_1 F \equiv F_z(\tau)\Pi_1 z + F_y(\tau)\Pi_1 y + G_1(\tau), \\ \frac{d\Pi_1 y}{d\tau} &= \Pi_0 f \equiv f(\bar{z}_0(0) + \Pi_0 z, \bar{y}_0(0) + \Pi_0 y, 0) \\ &\quad - f(\bar{z}_0(0), \bar{y}_0(0), 0), \end{aligned} \tag{8.23}$$

其中

$$\begin{aligned} G_1(\tau) &= (F_z(\tau) - \bar{F}_z(0))(\bar{z}_0'(0)\tau + \bar{z}_1(0)) + (F_y(\tau) \\ &\quad - \bar{F}_y(0))(\bar{y}_0'(0)\tau + \bar{y}_1(0)) + (F_t(\tau) - \bar{F}_t(0))\tau. \end{aligned} \tag{8.24}$$

一般对于分解式中足标为 k ($k = 1, 2, \cdots$) 的项有方程

$$\begin{aligned} \frac{d\bar{z}_{k-1}}{dt} &= \bar{F}_k \equiv \bar{F}_z(t)\bar{z}_k + \bar{F}_y(t)\bar{y}_k + F_k(t), \\ \frac{d\bar{y}_k}{dt} &= \bar{f}_k \equiv \bar{f}_z(t)\bar{z}_k + \bar{f}_y(t)\bar{y}_k + f_k(t), \end{aligned} \tag{8.25}$$

$$\frac{d\Pi_k z}{d\tau} = \Pi_k F \equiv F_z(\tau)\Pi_k z + F_y(\tau)\Pi_k y + G_k(\tau),$$
$$\frac{d\Pi_k y}{d\tau} = \Pi_{k-1} f. \tag{8.26}$$

这里 $f_k(t)$ 和 $F_k(t)$ 由 $\bar{z}_i(t)$, $\bar{y}_i(t)$ $(i = 0, 1, \cdots, k-1)$ 按确定的表达式计算. $\Pi_{k-1}f$ 是类似于 (8.18) 的 Πf 的分解式中 μ^{k-1} 的系数,它由 $\Pi_i z$, $\Pi_i y$ $(i = 0, 1, \cdots, k-1)$ 来表示.

为了由上述得到的方程来确定分解式 (8.14)、(8.15) 中的项,需要给出初始值. 应用初始条件 (8.2), 由 (8.13) 得

$$\bar{z}_0(0) + \mu\bar{z}_1(0) + \cdots + \Pi_0 z(0) + \mu\Pi_1 z(0) + \cdots = z_0,$$
$$\bar{y}_0(0) + \mu\bar{y}_1(0) + \cdots + \Pi_0 y(0) + \mu\Pi_1 y(0) + \cdots = y_0. \tag{8.27}$$

将 (8.27) 等号两边的同次幂系数相等,可以确定所需要的初始条件. 对于零次幂,我们得到

$$\bar{z}_0(0) + \Pi_0 z(0) = z_0, \quad \bar{y}_0(0) + \Pi_0 y(0) = y_0. \tag{8.28}$$

因为 $\bar{z}_0(t)$, $\bar{y}_0(t)$ 是 (8.20) 的解, 所以对于 $\bar{z}_0(t)$ 不需要给出附加的条件,而对于 $\bar{y}_0(t)$ 的自然的初始条件为

$$\bar{y}_0(0) = \bar{y}_0. \tag{8.29}$$

于是 (8.20)、(8.29) 的解将与退化组 (8.3)、(8.4) 的解 $\bar{z}(t) = \varphi(\bar{y}(t), t)$, $\bar{y}(t)$ 完全重合. 这样, 级数 (8.14) 的主项就是退化解.

考虑到 (8.29),由 (8.28) 得到 (8.21) 的初始条件为

$$\Pi_0 y(0) = 0, \tag{8.30}$$

$$\Pi_0 z(0) = z_0 - \bar{z}_0(0). \tag{8.31}$$

由 (8.30) 和 (8.21) 的第二个方程推得

$$\Pi_0 y(\tau) \equiv 0, \quad \tau \geqslant 0.$$

这样,考虑到 (8.29), 对于 $\Pi_0 z(\tau)$ 得到

$$\frac{d\Pi_0 z}{d\tau} = F(\bar{z}_0(0) + \Pi_0 z, y_0, 0). \tag{8.32}$$

这里 $\bar{z}_0(0) = \varphi(y_0, 0)$. 不难看出, (8.32) 可以由辅助组 (8.7)

通过代换 $\tilde{z} = \bar{z}_0(0) + \Pi_0 z$ 得到．因此(8.32)的静止点是 $\Pi_0 z = 0$．由于稳定性条件 (8.11)，静止点 $\Pi_0 z = 0$ 是渐近稳定的．由条件 V，初始条件 $\Pi_0 z(0) = z_0 - \bar{z}_0(0)$ 属于这个静止点的吸引区域，则当 $\tau \to \infty$ 时

$$\Pi_0 z(\tau) \to 0. \tag{8.33}$$

在［114］中给出了对于边界函数 $\Pi_i x(\tau)$ $(i = 0, 1, \cdots, n)$ 存在常数 $c > 0$ 和 $\kappa > 0$ 使成立不等式

$$\|\Pi_i x(\tau)\| \leqslant c \exp(-\kappa\tau),\ \tau \geqslant 0, \tag{8.34}$$

其中 κ 可取满足 $0 < \kappa < \alpha$ 的任意固定的数．因此条件 (8.33) 及下面的条件 (8.40) 均成立．这样，零次近似就完全确定了．

将 (8.23) μ 的一次幂项相等，得到

$$\bar{z}_1(0) + \Pi_1 z(0) = 0,\ \bar{y}_1(0) + \Pi_1 y(0) = 0. \tag{8.35}$$

首先考虑第二个等式．由于 (8.27) 的第二个方程，我们有

$$\Pi_1 y(\tau) = \Pi_1 y(0) + \int_0^\tau \Pi_0 f(s) ds.$$

由于我们要求当 $\tau \to \infty$ 时，边界函数趋于零，应该取

$$\Pi_1 y(0) = -\int_0^\infty \Pi_0 f(s) ds. \tag{8.36}$$

于是可得 $\Pi_1 y(\tau)$ 为

$$\Pi_1 y(\tau) = -\int_\tau^\infty \Pi_0 f(s) ds. \tag{8.37}$$

这样，$\bar{y}_1(0)$ 应该取成

$$\bar{y}_1(0) = \int_0^\infty \Pi_0 f(s) ds. \tag{8.38}$$

现在转向组 (8.22)．由于 $\mathrm{Re}\bar{\lambda}_i(t) < 0$，矩阵 $\bar{F}_z(t)$ 是非奇异的，由 (8.22) 的第一个方程可以解出 $\bar{z}_1(t)$，让它由 $\bar{y}_1(t)$ 和 $\frac{d\bar{z}_0}{dt}$ 来表示．将 $z_1(t)$ 的表达式代入 (8.22) 的第二个方程，得到 $\bar{y}_1(t)$ 的线性方程组．利用初始条件 (8.38) 可解出 $\bar{y}_1(t)$．这样得到 $\bar{z}_1(t)$，$\bar{y}_1(t)$．由 (8.35) 的第一个方程，得到 (8.23) 的第一个方程的初始条件

$$\Pi_1 z(0) = -\bar{z}_1(0), \tag{8.39}$$

再解这个方程可以确定 $\Pi_1 z(\tau)$，其中 $\Pi_1 y(\tau)$ 看成是已知.

完全类似地，应用附加条件

$$\Pi_k y(\tau) \to 0, \text{ 当 } \tau \to \infty, \tag{8.40}$$

$$\bar{y}_k(0) = \int_0^\infty \Pi_{k-1} f(s) ds, \tag{8.41}$$

$$\Pi_k z(0) = -\bar{z}_k(0). \tag{8.42}$$

可以由(8.25)、(8.26)确定 $\Pi_k y(\tau)$，$\bar{y}_k(t)$，$\bar{z}_k(t)$，$\Pi_k z(\tau)$ ($k=2, 3, \cdots$).

这样我们就得到初值问题(8.1)和(8.2)的解的分解式(8.13)—(8.15). 令 $x_n(t, \mu)$ 是分解式(8.13)的部分和

$$x_n(t, \mu) = \sum_{k=0}^{n} \mu^k [\bar{x}_k(t) + \Pi_k x(\tau)], \tag{8.43}$$

Васильева 证明了下面的定理:

定理 8.2 (Васильева) 设条件 I′、II、III、IV′、V 成立，其中条件 I′ 中只要求有到 $n+2$ 阶的连续偏导数，则可找到常数 $\mu_0 > 0$ 和 $c > 0$，使当 $0 < \mu \leqslant \mu_0$ 时，问题(8.1)和(8.2)的解 $z(t, \mu)$，$y(t, \mu)$ 在区间 $0 \leqslant t \leqslant T$ 上存在唯一，并满足估计式

$$\|x(t, \mu) - x_n(t, \mu)\| \leqslant c\mu^{n+1}, \quad 0 \leqslant t \leqslant T. \tag{8.44}$$

定理的证明见[114]，其中还给出(8.38)和(8.41)右边的积分的收敛性证明.

§2 边界层型数值方法

在§1中形式地给出了问题(8.1)、(8.2)的解的渐近展开式. 我们在这一节将描述一个数值方法，使能通过数值计算来求解确定形式展开中各个项的方程.

令 $h > 0$ 是变量 t 的离散步长. 记 $W = \begin{pmatrix} y \\ z \end{pmatrix}$ 和 $\Pi W = \begin{pmatrix} \Pi y \\ \Pi z \end{pmatrix}$，则它们均是 $M + m$ 维向量. 由定理 8.2 和(8.13)—(8.15)，

有

$$W(h, \mu) = \overline{W}_0(h) + \mu\overline{W}_1(h) + \Pi_0W(h/\mu) + \mu\Pi_1W(h/\mu) + O(\mu^2). \tag{8.45}$$

假定$\mu > 0$很小，步长h满足

$$h \gg \mu, \tag{8.46}$$

并与条件(8.33)，(8.40)一起推出 $\Pi_0W(h/\mu)$ 和 $\Pi_1W(h/\mu)$ 近似为零．因此，我们用 $\overline{W}_0(h) + \mu\overline{W}_1(h)$ 来近似 $W(h)$．近似解的误差是 $O(\mu^2)$（即随着μ的减小，它的精度提高）．数值方法是要计算 $\overline{W}_0(h)$ 和 $\overline{W}_1(h)$，但是仍必须计算 Π_0W，以便得到初始条件 $\bar{y}_1(0)$，这在确定 $\overline{W}_1(h)$ 时是需要的（如果需要还可以计算展开式中更多的项）．

数值方法由下面的步 1—步 4 所组成．

步 1 数值求解方程

$$\begin{aligned} \frac{d\bar{y}_0}{dt} &= f(\bar{z}_0, \bar{y}_0, t), \quad \bar{y}_0(0) = y_0, \\ 0 &= F(\bar{z}_0, \bar{y}_0, t), \end{aligned} \tag{8.47}$$

来确定 $\bar{y}_0(h)$，$\bar{z}_0(0)$ 和 $\bar{z}_0(h)$．

步 2 利用步 1 中确定的 $\bar{z}_0(0)$，数值求解初值问题

$$\frac{d\Pi_0z}{d\tau} = F(\bar{z}_0(0) + \Pi_0z, y_0, 0), \quad \Pi_0z(0) = z_0 - \bar{z}_0(0), \tag{8.48}$$

它是在变量τ的离散网格上进行的，得到解序列 $\Pi_0z(jh_\tau)$，$j=0, 1, \cdots, N$，其中步长h_τ和N的取法使得步 3 中的计算满足预先指定的精度．

步 3 用求积公式计算初值

$$\begin{aligned} \bar{y}_1(0) &= -\Pi_1y(0) = \int_0^\infty \Pi_0f(s)ds \\ &= \int_0^\infty [f(\bar{z}_0(0) + \Pi_0z(s), \bar{y}_0(0), 0) \\ &\quad - f(\bar{z}_0(0), \bar{y}_0(0), 0)]ds, \end{aligned} \tag{8.49}$$

得

$$\xi_1 = \sum_{j=0}^{N} a_j \Pi_0 f(jh_\tau)$$

$$= \sum_{j=0}^{N} a_j [f(\bar{z}_0(0) + \Pi_0 z(jh_\tau), \bar{y}_0(0), 0) - f(\bar{z}_0(0), \bar{y}_0(0), 0)], \tag{8.50}$$

其中 a_j 是求积公式的系数.

步 4 利用步 3 确定的 $\bar{y}_1(0)$ 的近似值 ξ_1, 求解初值问题

$$\frac{d\bar{y}_1}{dt} = \bar{f}_z(t)\bar{z}_1 + \bar{f}_y(t)\bar{y}_1, \quad \bar{y}_1(0) = \xi_1, \tag{8.51}$$

其中

$$\bar{z}_1(t) = -\bar{F}_z(t)^{-1}\left[\bar{F}_y(t)\bar{y}_1(t) - \frac{d\bar{z}_0(t)}{dt}\right]. \tag{8.52}$$

矩阵 $\bar{F}_z$、$\bar{F}_y$、$\bar{f}_z$ 和 $\bar{f}_y$ 仍按 § 1 中定义. 得到 $\bar{y}_1(h)$ 和 $\bar{z}_1(h)$.

得到 $\bar{y}_0(h)$, $\bar{y}_1(h)$ 和 $\bar{z}_0(h)$, $\bar{z}_1(h)$ 启, 可以构造问题 (8.1)、(8.2) 的解 $z(t, \mu)$, $y(t, \mu)$ 在点 $t_1 = h$ 处的近似 z_1, y_1 为

$$\begin{aligned} z_1 &= \bar{z}_0(h) + \mu\bar{z}_1(h), \\ y_1 &= \bar{y}_0(h) + \mu\bar{y}_1(h). \end{aligned} \tag{8.53}$$

再用 $t = h$ 代替步 1—步 4 中的 $t = 0$ 和用 z_1, y_1 代替其中的 z_0, y_0, 可以得到 $t = 2h, 3h, \cdots$ 上的近似解 z_2, y_2, z_3, y_3, $\cdots$.

在步 1 和步 4 中分别确定 $\bar{W}_0(h)$ 和 $\bar{W}_1(h)$, 而在步 2 和步 3 处理 $\Pi_0 z$ 并用来确定 $\bar{y}_1$ 的初始条件. 在这一部分中的计算精度可以比步 1 中的精度差一点. 事实上, 这一部分的计算误差将引起 $\bar{y}_1(0)$ 的一个误差. 从而在 $\bar{z}_1(h)$ 和 $\bar{y}_1(h)$ 引进一个相应的误差. 但是由于 (8.53) 和参数 μ, 这个误差对于 z_1, y_1 的影响将至少减小一个量级. 因此, μ 愈小, 步 2, 步 3 和步 4 的计算精度可愈差.

例 8.1 考虑初值问题

$$\frac{dy}{dt} = z - x, \quad y(0) = \xi, \tag{8.54}$$

$$\mu\frac{dz}{dt} = -z + \frac{1}{100}, \quad z(0) = \eta,$$

其中 $\mu=\frac{1}{100}$. 它的精确解为

$$y=\frac{1}{100}+\left(\xi+\frac{\eta-\frac{1}{100}}{99}-\frac{1}{100}\right)e^{-t}-\frac{\eta-\frac{1}{100}}{99}e^{-100t},$$

$$z=\frac{1}{100}+\left(\eta-\frac{1}{100}\right)e^{-100t}.$$

现在用前面叙述过的算法来求解这个问题.

(i) 解问题

$$\begin{aligned}&\frac{d\bar{y}_0}{dt}=\bar{z}_0-\bar{y}_0,\ \bar{y}_0(0)=\xi,\\&0=-\bar{z}_0+\frac{1}{100},\end{aligned}\tag{8.55}$$

得 $\bar{y}_0(h)$, $\bar{z}_0(0)$ 和 $\bar{z}_0(h)$. 用 t 的步长 h 解(8.55), 得

$$\begin{aligned}&\bar{y}_0(h)=\frac{1}{100}h+(1-h)\xi,\\&\bar{z}_0(0)=\bar{z}_0(h)=\frac{1}{100}.\end{aligned}\tag{8.56}$$

(ii) 在节点 $\tau_j=jh_\tau$, $j=0,1,\cdots,N$ 上求解

$$\frac{d\Pi_0z}{d\tau}=-\Pi_0z,\ \Pi_0z=\eta-\bar{z}_0(0)=\eta-\frac{1}{100}.\tag{8.57}$$

然后计算

$$\bar{y}_1(0)=\int_0^\infty\Pi_0z(s)ds\tag{8.58}$$

的近似值. 在(8.57)中用 τ 的步长 h_τ 的 Euler 方法, 并对(8.58)用矩形公式, 将上限用 Nh_τ 代替, 我们得到

$$\bar{y}_1(0)\approx\left(\eta-\frac{1}{100}\right)(1-(1-h_\tau)^N).\tag{8.59}$$

(iii) 求解问题

$$\begin{aligned}&\frac{d\bar{y}_1}{dt}=-\bar{y}_1,\quad\bar{y}_1(0)=\left(\eta-\frac{1}{100}\right)(1-(1-h_\tau)^N),\\&\bar{z}_1=0,\end{aligned}\tag{8.60}$$

应用步长为 h 的 Euler 方法，我们得到

$$\bar{y}_1(h) = (1-h)\left(\eta - \frac{1}{100}\right)(1-(1-h_\tau)^N),$$
$$\bar{z}_1(h) = 0. \tag{8.61}$$

再由 (8.56)，我们得到

$$y(h) \approx \frac{1}{100}h + (1-h)\xi + \frac{1}{100}(1-h)$$
$$\cdot\left(\eta - \frac{1}{100}\right)(1-(1-h_\tau)^N), \tag{8.62}$$
$$z(h) \approx \frac{1}{100}.$$

经整理后，(8.62) 的第一个式子为

$$y(h) \approx \frac{1}{100} + (1-h)\left(-\frac{1}{100} + \xi\right.$$
$$\left. + \frac{1}{100}\left(\eta - \frac{1}{100}\right)(1-(1-h_\tau)^N)\right).$$

上面给出的算法明显地依赖于小参数 μ. 下面推导一个不明显地依赖于小参数的数值方法.

记 $H = \begin{pmatrix} f \\ F \end{pmatrix}$, $W = \begin{pmatrix} y \\ z \end{pmatrix}$, 我们从给定的初值问题

$$\frac{dW}{dt} = H(t, W, \mu), \quad W(0) = \xi = \begin{pmatrix} y_0 \\ z_0 \end{pmatrix} \tag{8.63}$$

出发. 其中虽然 μ 已表示出来，但认为是尚未识别的. 以步长 h 沿 t 的节点数值求解 (8.63). 开始我们认为 (8.63) 不是刚性组，并且认为 W 的维数 N 即是 y 的维数 m，而 $M = 0$. 应用通常自开始的方法得到的 $t = h$ 的解，记之为 $W_0(h)$. 然后按分量来比较 $W_0(h)$ 和 ξ，即检验下面的不等式

$$\frac{|W_{0,i}(h) - \xi_i|}{1 + |\xi_i|} > \theta, \quad i = 1, \cdots, N, \tag{8.64}$$

其中 θ 是给定的一个小量. 如果 $W_0(h)$ 的分量均使 (8.64) 中的大于号不成立，我们就接受这一步产生的 $W_0(h)$ 的值. 如果对

于M个足标的组 J 中的足标 i，(8.64) 成立大于号，那么就拒绝这个积分步，按下面的方式重做.

令

$$y_i = W_i,$$
$$y_i^0 = \xi_i,$$
$$f_i = H_i, \ i = 1, \cdots, N, \ i \notin J.$$

和令

$$z_j = W_j,$$
$$z_j^0 = \xi_j,$$
$$F_j = H_j, \ j = 1, \cdots, N, \ j \in J.$$

现在 (8.63) 有形式

$$\begin{aligned} \frac{dy}{dt} &= f(t, y, z, \mu), \quad y(0) = y_0, \\ \frac{dz}{dt} &= F(t, y, z, \mu), \quad z(0) = z_0, \end{aligned} \tag{8.65}$$

其中参数μ仍未识别的. 但对 (8.65) 的右函数作下面的假定:

假定 8.1. 在 $\mu = 0$ 的邻域中，$f(t, y, z, \mu)$ 和 $F(t, y, z, \mu)$ 对μ均是解析的，而 $F(t, y, z, \mu)$ 在 $\mu = 0$ 处有一个单极点. 矩阵 $F_z(t, y, z, \mu)$ 的特征值的实部小于零.

我们寻找形式为

$$\begin{aligned} y(t) &= \bar{y}_0(t) + \mu\bar{y}_1(t) + \Pi_0 y(\tau) + \mu\Pi_1 y(\tau) + \cdots, \\ z(t) &= \bar{z}_0(t) + \mu\bar{z}_1(t) + \Pi_0 z(\tau) + \mu\Pi_1 z(\tau) + \cdots, \end{aligned} \tag{8.66}$$

的解. 对于外部解，要求它们满足

$$\frac{d\bar{y}_0}{dt} + \mu\frac{d\bar{y}_1}{dt} = f(t, \bar{y}_0, \bar{z}_0; \mu) + \mu f_y(t, \bar{y}_0, \bar{z}_0; \mu)\bar{y}_1 + \mu f_z(t, \bar{y}_0, \bar{z}_0; \mu)\bar{z}_1 + \cdots, \tag{8.67}$$

$$\frac{d\bar{z}_0}{dt} + \mu\frac{d\bar{z}_1}{dt} = F(t, \bar{y}_0, \bar{z}_0; \mu) + \mu F_y(t, \bar{y}_0, \bar{z}_0; \mu)\bar{y}_1 + \mu F_z(t, \bar{y}_0, \bar{z}_0; \mu)\bar{z}_1 + \cdots. \tag{8.68}$$

由假定，项 F, F_y和 F_z在 $\mu = 0$ 处有单极点. 因此，由 (8.67)、

(8.68) 对 $\bar{y}_0$, $\bar{z}_0$ 和 $\mu\bar{y}_1$, $\mu\bar{z}_1$ 分别导出下面的方程 (8.69) 和 (8.70).

$$
\begin{aligned}
\frac{d\bar{y}_0}{dt} &= f(t, \bar{y}_0, \bar{z}_0; \mu), \ \bar{y}_0(0) = y_0, \\
0 &= F(t, \bar{y}_0, \bar{z}_0; \mu).
\end{aligned}
\tag{8.69}
$$

注意这里没有假定 $\mu = 0$.

为简洁起见,省掉 f 和 F 的变量 $(t, \bar{y}_0, \bar{z}_0; \mu)$, 于是 $\mu\bar{y}_1$ 和 $\mu\bar{z}_1$ 的方程是

$$
\begin{aligned}
\frac{d\mu\bar{y}_1}{dt} &= f_y\mu\bar{y}_1 + f_z\mu\bar{z}_1, \\
\frac{d\bar{z}_0}{dt} &= F_y\mu\bar{y}_1 + F_z\mu\bar{z}_1.
\end{aligned}
\tag{8.70}
$$

由 (8.70) 的第二个方程解出 $\mu\bar{z}_1$, 并利用 (8.69), 得

$$
\begin{aligned}
\mu\bar{z}_1 &= F_z^{-1}\left[\frac{d\bar{z}_0}{dt} - F_y\mu\bar{y}_1\right] \\
&= F_z^{-1}[-F_z^{-1}(F_t + F_y f) - F_y\mu\bar{y}_1].
\end{aligned}
\tag{8.71}
$$

将 (8.71) 代入 (8.70), 得到确定 $\mu\bar{y}_1$ 和 $\mu\bar{z}_1$ 的方程

$$
\begin{aligned}
\frac{d\mu\bar{y}_1}{dt} &= (f_y - f_z F_z^{-1}F_y)\mu\bar{y}_1 - f_z F_z^{-2}(F_t + F_y f), \\
\mu\bar{z}_1 &= -F_z^{-1}F_y\mu\bar{y}_1 - F_z^{-2}(F_t + F_y f).
\end{aligned}
\tag{8.72}
$$

这里 μ 仍未确定,但要寻找的量 $\mu\bar{y}_1$ 和 $\mu\bar{z}_1$ 除初始条件 $\mu\bar{y}_1(0)$ 外均是确定的. 另外由假定 8.1, F, F_y, F_z 均是大的量. 在 (8.72) 中它们是以商的形式出现的. 在这个意义上大的量抵消掉了.

为了确定初始条件 $\mu\bar{y}_1(0)$, 将 (8.66) 代入 (8.65)

$$
\begin{aligned}
&\mu\bar{y}_0'(\mu\tau) + \mu^2\bar{y}_1'(\mu\tau) + \Pi_0'y(\tau) + \mu\Pi_1'y(\tau) + \cdots \\
&\quad = \mu f(\mu\tau; \bar{y}_0(\mu\tau) + \mu\bar{y}_1(\mu\tau) + \Pi_0 y(\tau) + \mu\Pi_1 y(\tau) \\
&\qquad + \cdots, \bar{z}_0(\mu\tau) + \cdots, \mu), \\
&\mu\bar{z}_0'(\mu\tau) + \mu^2\bar{z}_1'(\mu\tau) + \Pi_0'z(\tau) + \mu\Pi_1'z(\tau) + \cdots \\
&\quad = \mu F(\mu\tau, \bar{y}_0(\mu\tau) + \mu\bar{y}_1(\mu\tau) + \Pi_0 y(\tau) + \mu\Pi_1 y(\tau) \\
&\qquad + \cdots, \bar{z}_0(\mu\tau) + \cdots, \mu).
\end{aligned}
\tag{8.73}
$$

这里和下面，我们用 $'$ 表示对变量的微分.

应用假定 8.1，由 (8.73) 我们导出对 $\Pi_0 y$, $\Pi_1 y$ 和 $\Pi_0 z$ 的方程. 首先有

$$\frac{d\Pi_0 y}{d\tau} = 0, \tag{8.74}$$

由 $\Pi_0 y(0) + \bar{y}_0(0) = y_0$ 推出 $\Pi_0 y(0) = 0$，因此，$\Pi_0 y(\tau) = 0$. 对 $\Pi_1 y$ 和 $\Pi_0 z$ 的方程为

$$\frac{d\Pi_1 y}{d\tau} = f(0, y_0, \bar{z}_0(0) + \Pi_0 z(\tau); \mu) - f(0, y_0, \bar{z}_0(0); \mu), \tag{8.75}$$

$$\frac{d\Pi_0 z}{d\tau} = \mu F(0, y_0, \bar{z}_0(0) + \Pi_0 z(\tau); \mu). \tag{8.76}$$

应用边界层性质 $\lim\limits_{\tau\to\infty} \Pi y(\tau) = 0$，由零到$\infty$积分 (8.75)，再应用 $\bar{y}_1(0) + \Pi_1 y(0) = 0$，得到

$$\mu\bar{y}_1(0) = \mu \int_0^\infty [f(0, y_0, \bar{z}_0(0) + \Pi_0 z(\tau); \mu) - f(0, y_0, \bar{z}_0(0); \mu)] d\tau. \tag{8.77}$$

由于 $\Pi_0 z(\tau)$ 当 τ 从零增加时以指数的速度衰减，积分值的大部分是由 $\tau = 0$ 的邻域中得到. 因此用 $\tau = 0$ 的数据的插值来代替被积函数会得到积分的好的近似值. 首先，由初始条件 $\bar{z}_0(0) + \Pi_0 z(0) = z_0$，得到

$$\Pi_0 z(0) = z_0 - \bar{z}_0(0), \tag{8.78}$$

而由 (8.76) 本身，我们有

$$\Pi_0' z(0) = \mu F(0, y_0, z_0; \mu). \tag{8.79}$$

通过对 (8.76) 的微分，我们可以得到更多的数据，但下面我们只用 (8.78) 和 (8.79) 来近似. 最简单的近似是用在 $\tau = 0$ 处的切线的方程代替 (8.77) 中的被积函数，并且从零开始积分这切线的方程到它的正根. 按这种方式且由 (8.77)，得到

$$\mu\bar{y}_1(0) \approx -\frac{1}{2} \frac{[f(0, y_0, z_0; \mu) - f(0, y_0, \bar{z}_0(0); \mu)]^2}{f_z(0, y_0, z_0; \mu) F(0, y_0, z_0; \mu)}. \tag{8.80}$$

事实上，(8.77)的被积函数由线性函数

$$f(0, y_0, z_0; \mu) - f(0, y_0, \bar{z}_0(0); \mu) + \tau f_z(0, y_0, z_0; \mu)\mu F(0, y_0, z_0; \mu) \quad (8.81)$$

所近似. (8.81)的正根为

$$\tau_0 = -\frac{[f(0, y_0, z_0; \mu) - f(0, y_0, \bar{z}_0(0); \mu)]}{\mu f_z(0, y_0, z_0; \mu)F(0, y_0, z_0; \mu)}.$$

直接将(8.81)从零积分到 τ_0 再乘上 μ 即可得(8.80)式. 在(8.80)中的运算均是按分量进行的. 在分母中矩阵与向量相乘以后也取相应的分量.

为得到(8.77)的近似值也可采用指数拟合的方法. 令(8.77)的被积函数有形状 $ae^{b\tau}$. 当 $\tau = 0$ 时，有

$$a = f(0, y_0, z_0; \mu) - f(0, y_0, \bar{z}_0(0); \mu).$$

对 τ 微分，并令 $\tau = 0$，得

$$\frac{d}{d\tau}(ae^{b\tau})|_{\tau=0} = ab = \mu f_z(0, y_0, z_0; \mu)F(0, y_0, \bar{z}_0(0); \mu),$$

从而有

$$b = \frac{1}{a}[f_z(0, y_0, z_0; \mu)F(0, y_0, \bar{z}_0(0); \mu)].$$

通过直接积分可得

$$\mu\bar{y}_1(0) \approx \mu\frac{a}{b}e^{b\tau}|_0^\infty = -\mu\frac{a}{b}$$

$$= -\frac{[f(0, y_0, z_0; \mu) - f(0, y_0, \bar{z}_0(0); \mu)]^2}{f_z(0, y_0, z_0; \mu)F(0, y_0, z_0; \mu)}. \quad (8.82)$$

由(8.80)或(8.82)，以及(8.72)将完全确定 $\mu\bar{y}_1$ 和 $\mu\bar{z}_1$. 于是解(8.69)我们得到 $\bar{z}_0(0)$、$\bar{z}_0(h)$ 和 $\bar{y}_0(h)$. 再由(8.72)和(8.80)或(8.82)求出 $\mu\bar{y}_1(h)$ 和 $\mu\bar{z}_1(h)$，取

$$W(h) = \begin{pmatrix}\bar{y}_0(h) + \mu\bar{y}_1(h)\\ \bar{z}_0(h) + \mu\bar{z}_1(h)\end{pmatrix}.$$

现在我们在区间 $(h, 2h)$ 上重复上面的程序. 这时我们从已分成的(8.65)的组开始，作误差检验，将 $W(2h)$ 与 $W(h)$ 相比

较．若 $W(2h)$ 的所有分量均不使 (8.64) 中的大于号成立，则接受这一步；否则拒绝这一步，并按上面描述的格式分组及重做这一步．在重新分组时，已收入 z 组的分量将一直留在 z 组．

§3 渐近变换方法

§3.1 导数的拟稳定性

考虑在边界层外和边界层内确定刚性方程组的近似解的问题．如果已知微分方程组

$$\frac{dx}{dt}=f(t,x),\quad x(t)\in R_x^m \tag{8.83}$$

的一次积分，则向量 $x(t)$ 的一个分量（例如 $x^{(k)}(t)$）将可以由其余的分量及初值表示出来．将 $x^{(k)}(t)$ 的这个表示式代入除第 k 个以外的 (8.83) 的所有方程，我们得到阶数比原来的方程组少 1 的微分方程组．这个方程及代数方程一起与 (8.83) 是等价的．如果已知 (8.83) 的 k 个独立的一次积分，则由它我们可以导出等价的 $(m-k)$ 阶的微分方程组．在线性微分方程的情形，由已知的 k 个线性独立的特解也可以达到降阶的目的．但是在一般的情形寻找一次积分和在线性的情形寻找特解都是非常困难的．在实际计算中很少利用这种思想来构造算法．

但是，利用刚性微分方程组的特殊性，可以近似地寻找出其解向量的分量之间的代数关系式．例如，在前面已说明过，在边界层外，刚性线性方程确实建立了几乎精确的向量 $x(t)$ 的分量之间的代数关系式，关系式的个数等于这个组中快速衰减的特解的个数．因此，在边界层以外，原始方程组可用阶数较低的方程组来近似，并且比起原始组来，这个近似组的刚性比要小得多．下面想给出实际能行的方法来给出这个近似方程组．

考虑刚性线性方程组

$$\frac{dx}{dt}=Ax+b,\quad x(0)=x_0,\quad x(t)\in R_x^m, \tag{8.84}$$

将 (8.84) 改写成两个子组

$$\frac{dx_1}{dt} = A_{11}x_1 + A_{12}x_2 + b_1, \quad x_1(0) = x_{10}, \quad x_1(t) \in R^k_{x_1} \tag{8.85}$$

和

$$\frac{dx_2}{dt} = A_{21}x_1 + A_{22}x_2 + b_2, \quad x_2(0) = x_{20}, \quad x_2(t) \in R^{m-k}_{x_2}. \tag{8.86}$$

公式 (8.85) 和 (8.86) 中的向量 $x_1(t)$ 的坐标是 $x(t)$ 的一些分量. 这些分量的个数 k 等于 (8.84) 的齐次方程的在边界层外可以略去的线性独立特解的个数. 向量 $x(t)$ 的其余的分量组成向量 $x_2(t)$. 实际如何组成向量 $x_1(t)$ 将在后面来说明. 矩阵 A_{11}, A_{12}, A_{21}, A_{22} 是由 $x_1(t)$, $x_2(t)$ 的分量所对应的矩阵 A 中的元素所组成. 设矩阵 A 的特征值排成下面的次序

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_k| > |\lambda_{k+1}| \geqslant \cdots \geqslant |\lambda_m|,$$
$$|\lambda_{k+1}| \ll |\lambda_k|, \tag{8.87}$$

并且足标 $i \leqslant k$ 的 λ_i 有估计

$$\exp(\mathrm{Re}\lambda_i \tau_{BL}) \leqslant \frac{1}{N}, \quad N \gg 1, \tag{8.88}$$

其中 τ_{BL} 是边界层的时间.

将矩阵 A 变换成标准的 Jordan 型

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} = UAU^{-1}, \tag{8.89}$$

其中矩阵 Λ_1 的维数为 $k \times k$, 有特征值 $\lambda_1, \lambda_2, \cdots, \lambda_k$. 按下面的表示将矩阵 A, U, U^{-1} 分块成子矩阵

$$U = \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}. \tag{8.90}$$

矩阵 U_{11} 和 Q_{11} 的阶数为 $k \times k$, 并且假定是非奇异的.

作变量代换,引进 k 维向量 y_1

$$y_1 = U_{11}x_1 + U_{12}x_2, \tag{8.91}$$

并且消去分量 x_1, 得

$$\frac{dy_1}{dt} = U_{11}(A_{11} + U_{11}^{-1}U_{12}A_{21})U_{11}^{-1}y_1 + U_{11}(A_{12} + U_{11}^{-1}U_{12}A_{22}$$

$$- A_{11}U_{11}^{-1}U_{12} - U_{11}^{-1}U_{12}A_{21}U_{11}^{-1}U_{12})x_2 + U_{11}b_1 + U_{12}b_2, \tag{8.92}$$

$$\frac{dx_2}{dt} = A_{12}U_{11}^{-1}y_1 + (A_{22} - A_{21}U_{11}^{-1}U_{12})x_2 + b_2, \tag{8.93}$$

$$y_{10} = U_{11}x_{10} + U_{12}x_{20}.$$

根据表达式 (8.89)

$$U_{11}A_{11} + U_{12}A_{21} = \Lambda_1U_{11},$$

$$U_{11}A_{12} + U_{12}A_{22} = \Lambda_1U_{12},$$

则对于 (8.92) 中的矩阵系数，我们有

$$U_{11}(A_{11} + U_{11}^{-1}U_{12}A_{21})U_{11}^{-1} = \Lambda_1$$

$$A_{12} + U_{11}^{-1}U_{12}A_{22} - A_{11}U_{11}^{-1}U_{12} - U_{11}^{-1}U_{12}A_{21}U_{11}^{-1}U_{12} = 0 \tag{8.94}$$

并且方程 (8.92) 变换成形式

$$\frac{dy_1}{dt} = \Lambda_1y_1 + U_{11}b_1 + U_{12}b_2. \tag{8.95}$$

由于变换 (8.91),(8.92)—(8.93)的矩阵与矩阵 A 相似. 由于 Λ_1 的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_k$, 矩阵 $(A_{22} - A_{21}U_{11}^{-1}U_{12})$ 的特征值为 $\lambda_{k+1}, \cdots, \lambda_m$.

(8.95) 的解有形式

$$y_1(t) = e^{\Lambda_1 t}(y_{10} + \Lambda_1^{-1}U_{11}b_1 + \Lambda_1^{-1}U_{12}b_2)$$

$$+ (-\Lambda_1)^{-1}(U_{11}b_1 + U_{12}b_2).$$

在边界层外，由于不等式 (8.88)，上述式子右边的第一项可以看成是不存在的，得到

$$y_1(t)|_{t>\tau_{BL}} \approx \bar{y}_1(t) = -\Lambda_1^{-1}(U_{11}b_1 + U_{12}b_2). \tag{8.96}$$

为了得到边界层外的 $x_1(t)$, $x_2(t)$ 的近似值，将 (8.96) 代入 (8.91) 和 (8.93)，我们得到

$$U_{11}\bar{x}_1 + U_{12}\bar{x}_2 + \Lambda_1^{-1}(U_{11}b_1 + U_{12}b_2) = 0, \quad t \geqslant \tau_{BL}, \tag{8.97}$$

$$\frac{d\bar{x}_2}{dt} = (A_{22} - A_{21}U_{11}^{-1}U_{12})\bar{x}_2 + b_2$$

$$-A_{21}U_{11}^{-1}\Lambda_1^{-1}(U_{11}b_1+U_{12}b_2). \tag{8.98}$$

为了得到边界层外向量 $x(t)$ 的坐标之间的线性关系式 (8.97)，将方程 (8.85) 微分 $s-1$ 次，并令 $x_1(t)$ 的第 s 次导数为零. 在对 (8.85) 的每次微分后，将 $x_1(t)$ 和 $x_2(t)$ 的一阶导数分别换成 (8.85) 和 (8.86) 中的表达式. 这样，当 $t \geqslant \tau_{BL}$ 时，$x_1(t)$ 的 $(s-1)$ 次导数认为是拟稳定的. 下面来说明这种处理的合理性.

对于解向量的 s 阶导数，有

$$\frac{d^s x}{dt^s}=\frac{d^{s-1}}{dt^{s-1}}(Ax+b)=A^s x+A^{s-1}b.$$

由表达式 (8.89)、(8.90)，矩阵 A^s 写成形式

$$A^s=\begin{pmatrix}Q_{11} & Q_{12}\\ Q_{21} & Q_{22}\end{pmatrix}\begin{pmatrix}\Lambda_1^s & 0\\ 0 & \Lambda_2^s\end{pmatrix}\begin{pmatrix}U_{11} & U_{12}\\ U_{21} & U_{22}\end{pmatrix}$$
$$=\begin{pmatrix}Q_{11}\Lambda_1^s U_{11}+Q_{12}\Lambda_2^s U_{21} & Q_{11}\Lambda_1^s U_{12}+Q_{12}\Lambda_2^s U_{22}\\ Q_{21}\Lambda_1^s U_{11}+Q_{22}\Lambda_2^s U_{21} & Q_{21}\Lambda_1^s U_{12}+Q_{22}\Lambda_2^s U_{22}\end{pmatrix}.$$

对于向量 $x_1(t)$ 有

$$\frac{d^s x_1}{dt^s}=(Q_{11}\Lambda_1^s U_{11}+Q_{12}\Lambda_2^s U_{21})x_1+(Q_{11}\Lambda_1^s U_{12}$$
$$+Q_{12}\Lambda_2^s U_{22})x_2+(Q_{11}\Lambda_1^{s-1}U_{11}+Q_{12}\Lambda_2^{s-1}U_{21})b_1$$
$$+(Q_{11}\Lambda_1^{s-1}U_{12}+Q_{12}\Lambda_2^{s-1}U_{22})b_2, \tag{8.99}$$

形式上令向量 $d^s x_1/dt^s$ 为零. 这时的 $x_1(t)$ 和 $x_2(t)$ 分别用 $\tilde{x}_1(t)$ 和 $\tilde{x}_2(t)$ 表示，再用 $\Lambda_1^{-s}Q_{11}^{-1}$ 左乘 (8.99)，我们得到表示式

$$(U_{11}+\Lambda_1^{-s}Q_{11}^{-1}Q_{12}\Lambda_2^s U_{21})\tilde{x}_1(t)+(U_{12}+\Lambda_1^{-s}Q_{11}^{-1}Q_{12}\Lambda_2^s U_{22})\tilde{x}_2(t)$$
$$+\Lambda_1^{-1}[(U_{11}+\Lambda_1^{1-s}Q_{11}^{-1}Q_{12}\Lambda_2^{s-1}U_{21})b_1$$
$$+(U_{12}+\Lambda_1^{1-s}Q_{11}^{-1}Q_{12}\Lambda_2^{s-1}U_{22})b_2]=0. \tag{8.100}$$

如果选 s 的值使得 $\|\Lambda_1^{-s}\|\|\Lambda_2^s\|\ll 1$，则 (8.100) 的每一项的第二个加项都可以忽略掉，得到 (8.97). 这里 $\|\cdot\|$ 指任意采用的矩阵模. 公式 (8.97) 还可以改写成 $\bar{x}_1(t)$ 的显式形式

$$\bar{x}_1+U_{11}^{-1}U_{12}\bar{x}_2+U_{11}^{-1}\Lambda_1^{-1}U_{11}b_1+U_{11}^{-1}\Lambda_1^{-1}U_{12}b_2=0.$$

根据 (8.100)，其中向量 $\bar{x}_2$、b_1、b_2 的矩阵系数有近似表示式

$$U_{11}^{-1}U_{12} \approx D_s(Q_{11}\Lambda_1^s U_{12} + Q_{12}\Lambda_2^s U_{22}), \tag{8.101}$$

$$U_{11}^{-1}\Lambda_1^{-1}U_{11} \approx D_s(Q_{11}\Lambda_1^{s-1}U_{11} + Q_{12}\Lambda_2^{s-1}U_{21}), \tag{8.102}$$

$$U_{11}^{-1}\Lambda_1^{-1}U_{12} \approx D_s(Q_{11}\Lambda_1^{s-1}U_{12} + Q_{12}\Lambda_2^{s-1}U_{22}), \tag{8.103}$$

其中

$$D_s = (Q_{11}\Lambda_1^s U_{11} + Q_{12}\Lambda_2^s U_{21})^{-1}.$$

因此，为了寻找边界层外 (8.85) 和 (8.86) 的解，可以利用维数较小的微分方程组 (8.98) 和代数关系式 (8.97)，而为了实际确定这个关系式，要利用 $x_1(t)$ 的 $s-1$ 阶导数的拟稳定性条件.

在对刚性系统 (8.84) 进行近似时，导数的拟稳定性条件不仅可以应用到边界层外，也可以应用到边界层内. 如果从矩阵 A 的第 $k+1$ 个特征值开始，特征值都非常小. 即

$$|\lambda_k| \gg |\lambda_{k+1}|, \quad |\lambda_{k+1}|\tau_{BL} \ll 1.$$

则当 $t < \tau_{BL}$ 时，(8.84) 的解可以用 k 阶微分方程组和其余变量的幂级数表示式来精确描述，这个 k 阶组的矩阵的特征值为 $\lambda_1, \lambda_2, \cdots, \lambda_k$.

为了建立这样的组，在方程 (8.93) 中进行变量代换

$$x_2 = W_{21}U_{11}^{-1}y_1 + y_2, \tag{8.104}$$

如果选取矩阵 W_{21} 满足矩阵方程

$$(A_{22} - A_{21}U_{11}^{-1}U_{12})W_{21} + A_{21} = W_{21}(A_{11} + U_{11}^{-1}U_{12}A_{21}), \tag{8.105}$$

则由 (8.93) 得到的 $y_2(t)$ 不依赖于 $y_1(t)$:

$$\frac{dy_2}{dt} = (A_{22} - A_{21}U_{11}^{-1}U_{12})y_2 + b_2 - W_{21}(b_1 + U_{11}^{-1}U_{12}b_2), \tag{8.106}$$

$$y_2(0) = y_{20} = x_{20} - W_{21}(x_{10} + U_{11}^{-1}U_{12}x_{20}).$$

这样，(8.85) 和 (8.86) 的迅速衰减的解向量由 k 阶 (8.95) 来描述. 它可以写成

$$\frac{d(U_{11}^{-1}y_1)}{dt} = (A_{11} + U_{11}^{-1}U_{12}A_{21})(U_{11}^{-1}y_1) + b_1 + U_{11}^{-1}U_{12}b_2. \tag{8.107}$$

$$U_{11}^{-1}y_{10} = x_{10} + U_{11}^{-1}U_{12}x_{20}.$$

而为了描述边界层外的解，应用维数为 $(m-k)$ 的(8.106)。

根据(8.91)和(8.104)，对于 $x_1(t)$ 和 $x_2(t)$，我们有

$$x_1=(E-U_{11}^{-1}U_{12}W_{21})(U_{11}^{-1}y_1)-U_{11}^{-1}U_{12}y_2 \tag{8.108}$$

$$x_2=W_{21}U_{11}^{-1}y_1+y_2. \tag{8.109}$$

根据导数的拟稳定性原则，公式(8.106)—(8.109)中的矩阵 $U_{11}^{-1}U_{12}$ 由公式(8.101)得到.

为了得到矩阵 W_{21}，应用方程(8.105)，它有唯一解，这是因为矩阵 $(A_{11}+U_{11}^{-1}U_{12}A_{21})$ 和 $(A_{22}-A_{21}U_{11}^{-1}U_{12})$的特征值是不同的. 将(8.105)记成

$$W_{21}=[A_{21}+(A_{22}-A_{21}U_{11}^{-1}U_{12})W_{21}](A_{11}+U_{11}^{-1}U_{12}A_{21})^{-1},$$

并且应用简单迭代法来求解. 由于量 $|\lambda_k|^{-1}|\lambda_{k+1}|$ 非常小，用简单迭代法收敛是非常快的. 逆矩阵

$$(A_{11}+U_{11}^{-1}U_{12}A_{21})^{-1}=U_{11}^{-1}\Lambda_1^{-1}U_{11} \tag{8.110}$$

及 $U_{11}^{-1}U_{12}$ 可以根据导数的拟稳定性原则按公式(8.102)计算.

这样确定 W_{21} 的递推关系式有形式

$$W_{21}^{l+1}=[A_{21}+(A_{22}-A_{21}U_{11}^{-1}U_{12})W_{21}^{l}]U_{11}^{-1}\Lambda_1^{-1}U_{11},\quad W_{21}^{0}=0. \tag{8.111}$$

对于刚性系统(8.84)，因为 $|\lambda_k|^{-1}|\lambda_{k+1}|\ll 1$，公式(8.111)一般迭代一次就行.

这样，为了在边界层内描述过程，只要求解 k 阶(8.107)，而当 $t<\tau_{BL}$ 时，子向量 $y_2(t)$ 由 ν 不大的 Taylor 公式

$$y_2(t)=y_2(0)+\frac{t}{1!}y_2'(0)+\cdots+\frac{t^\nu}{\nu!}\frac{d^\nu y_2(0)}{dt^\nu}$$

来计算，其中系数根据(8.106)得到.

在边界层外，如果选取初始条件 $\bar{x}_2(\tau_{BL})=x_2(\tau_{BL})$，则刚性系统(8.84)可以由(8.97)和(8.98)来描述. 当 $t\geqslant\tau_{BL}$ 和特殊选取的初始条件

$$\begin{aligned}\bar{x}_2(0)=x_{20}-W_{21}[&x_{10}+U_{11}^{-1}U_{12}x_{20}\\&+U_{11}^{-1}\Lambda_1^{-1}U_{11}(b_1+U_{11}^{-1}U_{12}b_2)]\end{aligned} \tag{8.112}$$

时，这些公式也可以应用。(8.112)是由(8.106)—(8.109)推得

的．事实上，当 $t \geqslant \tau_{BL}$ 时，对于子向量 $\tilde{x}_2(t)$ 有

$$\tilde{x}_2(t) = -W_{21}(A_{11} + U_{11}^{-1}U_{12}A_{21})^{-1}(b_1 + U_{11}^{-1}U_{12}b_2) + y_2(t). \tag{8.113}$$

令 $t=0$，并将 (8.106) 中的 $y_2(0)$ 的值代入(8.113)，得到(8.112)．

下面给出当 $t \geqslant \tau^* \geqslant \tau_{BL}$ 时，应用导数的拟稳定性原则的误差的渐近估计．记

$$\begin{aligned} D_s &= (Q_{11}\Lambda_1^s U_{11} + Q_{12}\Lambda_2^s U_{21})^{-1}, \\ V_s &= D_s(Q_{11}\Lambda_1^s U_{12} + Q_{12}\Lambda_2^s U_{22}), \\ \varepsilon_1 &= x_1 - \tilde{x}_1,\ \varepsilon_2 = x_2 - \tilde{x}_2,\ \varepsilon = (\varepsilon_1, \varepsilon_2)^T. \end{aligned} \tag{8.114}$$

将公式 (8.85) 微分 $(s-1)$ 次，左乘 D_s，减去 (8.100)，再左乘 U_{11}^{-1}，得到方程

$$\varepsilon_1 + V_s\varepsilon_2 = D_s \frac{d^s x_1}{dt^s}. \tag{8.115}$$

类似地，由 (8.86) 减去方程组

$$\frac{d\tilde{x}_2}{dt} = A_{21}\tilde{x}_1 + A_{22}\tilde{x}_2 + b_2, \tag{8.116}$$

得到方程

$$\frac{d\varepsilon_2}{dt} = A_{21}\varepsilon_1 + A_{22}\varepsilon_2. \tag{8.117}$$

利用 (8.115) 消去 ε_1，得到

$$\frac{d\varepsilon_2}{dt} = (A_{22} - A_{21}V_s)\varepsilon_2 + A_{21}D_s \frac{d^s x_1}{dt^s}. \tag{8.118}$$

(8.118) 的初始条件取成

$$\varepsilon_2^* = \varepsilon_2(\tau^*),$$

$\tau^* \geqslant \tau_{BL}$．若记 $B_s = A_{22} - A_{21}V_s$，则 (8.118) 的解有形式

$$\begin{aligned} \varepsilon_2(t) &= \exp[B_s(t-\tau^*)]\varepsilon_2^* \\ &\quad + \int_{\tau^*}^{t} \exp[B_s(t-\tau)]A_{21}D_s \frac{d^s x_1(\tau)}{d\tau^s}\, d\tau. \end{aligned} \tag{8.119}$$

引进渐近参数

$$\mu = \|\Lambda_1^{-s}\| \|\Lambda_2^s\| \to 0, \tag{8.120}$$

则由于组 (8.85) 和 (8.86) 是刚性系统，有估计

$$\|A_{22}-A_{21}U_{11}^{-1}U_{12}-B_s\|\leqslant\|A_{21}\|\|V_s-U_{11}^{-1}U_{12}\|=O(\mu),\quad \mu\to 0,$$

另外，在边界层外，当 $t\geqslant\tau^*$ 时，

$$\|D_s\|\left\|\frac{d^s x_1(t)}{dt^s}\right\|=0(\mu),\quad \mu\to 0.$$

因此，若选取初始条件

$$\varepsilon_2^*=O(\mu),\quad \mu\to 0,$$

得到

$$\|\varepsilon_2(t)\|=O(\mu),\quad \mu\to 0.$$

总结上面所叙述的，得到下面的定理:

定理 8.3 (导数拟稳定性原则) 对于刚性微分方程组 (8.84) 和充分小的 $\varepsilon>0$，可以找到整数 $s(\varepsilon)\geqslant 1$，向量 $x(t)$ 的子向量 $x_1(t)$ 和数 $\tau^*=\tau(\varepsilon)$，$\tau_{BL}\leqslant\tau(\varepsilon)\leqslant T$，使得方程 (8.85) 和 (8.86) 的解满足不等式

$$\|x_i(t)-\tilde{x}(t)\|\leqslant\varepsilon\max_t\|x_i(t)\|,\quad i=1,2,;\ t\geqslant\tau(\varepsilon),$$

并且其中向量 $\tilde{x}_1(t)$ 是由 k 个代数关系式 (8.100) 确定，而向量 $\tilde{x}_2(t)$ 是由 $m-k$ 个微分方程 (8.116) 及初始向量 $\tilde{x}_2(\tau^*)=x_2(\tau^*)$ 确定.

在实际应用这个定理时，需要解决一系列问题. 首先一个问题是如何将 (8.84) 的微分方程进行分类，使其中的一部分放进 (8.116)，而将另一部分由代数关系式来处理.

进行分类的一个基本的准则是对于两个相邻的 s 值，矩阵 $B_s=A_{22}-A_{21}V_s$ 的元素的相合性，并且矩阵 B_s 的谱半径比起 (8.84) 的矩阵 A 的谱半径的要小得多，或者用 (8.100)、(8.116) 的解的相合性来验证.

虽然随着 s 的值增大 ε 的值将减小，但是由于初始组的矩阵是以有限精度给出的，而计算中要引进舍入误差，s 的值取得太大是不合理的，通常取 $s=2,3$. 如果按照模，矩阵 A 的特征值分成两个非常不同的类，并且 $|\lambda_{k+1}|\ll|\lambda_k|$，则根据 (8.120)，对于这

些 s 的值参数 μ 是充分小.

在选取(8.100)的方程时,有两种极限情形. 一种是刚性系统(8.84) 可以分出导数含有小参数的方程, 系统的运动分成快变的和慢变的. 那么为了得到代数关系式, 选取对应于快变变量的方程. 另外一种极限情形是不存在这种分法,所有的分量实际上是平等的,并且 A 的所有行均与按模的大特征值相关,则从理论上来看选取哪个方程到 (8.100) 中去都是一样的.

为了说明导数拟稳定性原则的应用,考虑下面的例子.

例 8.2. 考虑二阶系统

$$\frac{dx^{(1)}}{dt} = -501x^{(1)} + 500x^{(2)}, \quad x^{(1)}(0) = x_{10}; \tag{8.121}$$

$$\frac{dx^{(2)}}{dt} = 500x^{(1)} - 501x^{(2)}, \quad x^{(2)}(0) = x_{20}, \quad t \in [0, 1], \tag{8.122}$$

它有显式解

$$\begin{aligned} x^{(1)}(t) &= 0.5(x_{10} - x_{20})\exp(-1001t) + 0.5(x_{10} + x_{20})\exp(-t), \\ x^{(2)}(t) &= -0.5(x_{10} - x_{20})\exp(-1001t) + 0.5(x_{10} + x_{20})\exp(-t). \end{aligned} \tag{8.123}$$

显然,对于 $t \geqslant \tau_{BL}(\tau_{BL} \sim 10^{-2})$, 向量 $x(t)$ 的分量之间存在线性关系

$$\bar{x}^{(1)}(t) = \bar{x}^{(2)}(t).$$

为了用 $s = 1$ 的公式(8.100)和(8.116)来近似描述边界层外的系统,略去(8.121)中的一阶导数, 得

$$\tilde{x}^{(1)} = \frac{500}{501}\tilde{x}^{(2)},$$

$$\frac{dx^{(2)}}{dt} \approx -2\tilde{x}^{(2)}.$$

可以看出,它还不能保证精度. 这时

$U_{11}^{-1}U_{12} \approx -0.998$, $W_{21} \approx 0.5$, $A_{11} + U_{11}^{-1}U_{12}A_{21} \approx -1000$,

由于(8.108), (8.109)的近似系统的解的第二个指数有较大的误差,即有

$$\tilde{x}^{(1)} \simeq (0.501x_{10} - 0.500x_{20})\exp(-1000t)$$
$$+ (0.499x_{10} + 0.500x_{20})\exp(-2t),$$
$$\tilde{x}^{(2)} \simeq (-0.500x_{10} + 0.499x_{20})\exp(-1000t)$$
$$+ (0.500x_{10} + 0.501x_{20})\exp(-2t),$$

但是当 $s=2$ 时，有

$$\tilde{x}^{(1)} = \frac{500 \times 501 + 500 \times 501}{501 \times 501 + 500 \times 500}\tilde{x}^{(2)} \simeq 0.999998\tilde{x}^{(2)},$$

$$\frac{d\tilde{x}^{(2)}}{dt} \simeq -1.001\tilde{x}^{(2)},$$

这已经相当精确. 线性关系式中系数的误差阶为 2.10^{-6}，而指数的误差阶是 10^{-3}. 由公式 (8.108) 和 (8.109) 得到的解与 (8.123) 相比，可有 5 位有效数字.

要消去的方程的个数 k 应该对应于(8.84)的线性无关的快速衰减的特解的个数. 如果这个量预先不知道，则应该从不大的值 k 开始，只要变换得到的组的矩阵 B_s 的模不是所需要的值，就增大 k 的值. 而当 B_s 满足条件

$$\|B_s\| \leqslant L/N,$$

则停止在 $x_1(t)$ 的最小的维数上，其中 $N \gg 1$，而 L 是矩阵 A 的模 $\|A\|$ 的某种估计.

§ 3.2 非线性刚性系统导数的拟稳定性

导数的拟稳定性原则也可应用到非线性刚性系统 (8.83). 可以有两种方式，一种方式是假定在每一个时间子区间 $[t_i, t_i + H]$ 上，成立估计式

$$\left\|\frac{\partial f(t, x)}{\partial x} - A_i\right\| < \varepsilon, \quad t \in [t_i, t_i + H],$$

这里 A_i 是常矩阵，H 满足 $\tau_{BL_i} \ll H$. τ_{BL_i} 是对应于矩阵 A_i 的线性方程

$$\frac{dx}{dt} = A_i x$$

的边界层的时间．ε 是充分小的量．于是我们将整个积分区间分成子区间，而在每个子区间上用线性系统 (8.84) 来近似非线性系统 (8.83)，然后应用公式 (8.100) 和 (8.116)．

另一种方式是将导数的拟稳定性条件直接应用到 (8.1)，而得到边界层外 (8.83) 的近似．类似于 (8.85) 和 (8.86)，将方程组 (8.83) 改写成

$$\frac{dx_1}{dt}=f_1(t, x_1, x_2),\ x_1(t_0)=x_{10},\ x_1(t)\in R^k_{x_1}, \tag{8.124}$$

$$\frac{dx_2}{dt}=f_2(t, x_1, x_2),\ x_2(t_0)=x_{20},\ x_2(t)\in R^{m-k}_{x_2}, \tag{8.125}$$

$$t\in[t_0, t_0+T].$$

类似于 (8.100) 的代数关系式可以这样来得到：对于 (8.124) 微分 $s-1$ 次，令得到的结果为零，即

$$\phi(t, \tilde{x}_1, \tilde{x}_2)=\frac{d^{s-1}f_1(t, \tilde{x}_1, \tilde{x}_2)}{dt^{s-1}}=0, \tag{8.126}$$

并且在每次微分后将 $\frac{d\tilde{x}_1}{dt}$ 换成 $f_1(t, \tilde{x}_1, \tilde{x}_2)$，将 $d\tilde{x}_2/dt$ 换成 $f_2(t, \tilde{x}_1, \tilde{x}_2)$．假定 (8.126) 对 $\tilde{x}_1$ 是可解的，得到 $\tilde{x}_2(t)$ 的方程

$$\frac{d\tilde{x}_2(t)}{dt}=f_2(t, \tilde{x}_1, \tilde{x}_2),\ \tilde{x}_2(t_0+\tau^*)=x_2(t_0+\tau^*),\ \tau^*\geqslant\tau_{BL}. \tag{8.127}$$

进行的推导的正确性可通过检查不等式

$$\|\tilde{x}_1(t_0+\tau^*)-x_1(t_0+\tau^*)\|\leqslant\varepsilon\|x_1(t_0+\tau^*)\|$$

是否成立以及对相邻两个 s 值 (8.126) 和 (8.127) 的解是否重合来确定．并且要求变换后的 Jacobi 矩阵的谱半径应该小于原始的谱半径．关于向量 $\tilde{x}_1$ 及它的维数的选取，以及量 s 的大小均可以与线性的情形类似地解决．应该指出，为得到代数关系式 (8.126)，(8.124) 中方程的微分次数可以是不同的．

为了说明导数的拟稳定性原则应用到非线性方程的特殊性，考虑导数前具有小参数的微分方程

$$\mu \frac{dx}{dt} = f(t, x).$$

假定 $f(t, x)$ 是充分连续可微的．记

$$\mu \frac{d^s x}{dt^s} = \frac{d^{s-1} f(t, x)}{dt^{s-1}}, \tag{8.128}$$

$$0 = \frac{d^{s-1} f(t, \tilde{x})}{dt^{s-1}}, \tag{8.129}$$

并且估计边界层外差 $x(t) - \tilde{x}(t)$ 的量值．为此，将第一个方程减去第二个方程，并应用有限增量公式，有

$$\frac{\partial}{\partial x}\left[\frac{d^{s-1} f(t, x)}{dt^{s-1}}\right]_{x=x^*}(x - \tilde{x}) = \mu \frac{d^s x}{dt^s}. \tag{8.130}$$

这里值 x^* 在 $x(t)$ 和 $\tilde{x}(t)$ 之间．引进表示式

$$G_s(t, x, \mu) = \mu^{s-1} \frac{\partial}{\partial x}\left[\frac{d^{s-1} f(t, x)}{dt^{s-1}}\right]$$

并将 (8.130) 写成形式

$$G_s(t, x^*, \mu)(x - \tilde{x}) = \mu^s \frac{d^s x}{dt^s}.$$

如果在边界层外的求解区间上，量

$$\varepsilon_s = (x - \tilde{x}) = G_s^{-1}(t, x^*, \mu)\mu^s \frac{d^s x}{dt^s} \tag{8.131}$$

按模充分小，则当 $t \geqslant t_0 + \tau_{BL}$，(8.128) 可以用 (8.129) 来近似．

下面我们给出 $|\varepsilon_s|$ 为小的充分性条件．微分并应用数学归纳法，可以证明下面的公式

$$G_s(t, x, \mu) = \mu^{s-1} \frac{\partial}{\partial x}\left[\frac{d^{s-1} f(t, x)}{dt^{s-1}}\right]$$

$$= \left(\frac{\partial f}{\partial x}\right)^s + R_s(t, x, \mu),$$

其中 $R_s(t, x, \mu)$ 的每一个加项至少含有高于一阶 $f(t, x)$ 的偏导数作为因子，并且 $R_s(t, x, \mu)$ 对 μ 是连续的．事实上

$$G_{s+1}(t, x, \mu) = \mu^s \frac{\partial}{\partial x}\left[\frac{d^s f}{dt^s}\right]$$

$$= \mu^s \frac{\partial}{\partial x}\left[\frac{\partial}{\partial x}\left(\frac{d^{s-1}f}{dt^{s-1}}\right)\frac{f}{\mu} + \frac{\partial}{\partial t}\left(\frac{d^{s-1}f}{dt^{s-1}}\right)\right]$$

$$= \frac{\partial f}{\partial x} G_s(t, x, \mu) + \mu \frac{d}{dt} G_s(t, x, \mu)$$

$$= \left(\frac{\partial f}{\partial x}\right)^{s+1} + R_{s+1}(t, x, \mu),$$

其中

$$R_{s+1}(t, x, \mu) = \frac{\partial f}{\partial x} R_s(t, x, \mu) + \mu \frac{d}{dt}\left(\frac{\partial f}{\partial x}\right)^s + \mu \frac{dR_s(t, x, \mu)}{dt}.$$

例如，对于 $s = 2$

$$R_2(t, x, \mu) = \frac{\partial^2 f(t, x)}{\partial x^2} f(t, x) + \mu \frac{\partial^2 f(t, x)}{\partial t \partial x}.$$

现在假定在解的邻域中，函数 $\partial f(t, x)/\partial x$ 变化很小

$$\left|\frac{\partial f(t, x)}{\partial x}\right|^s \gg |R_s(t, x, \mu)|, \tag{8.132}$$

$$G_s(t, x, \mu) \approx \left(\frac{\partial f}{\partial x}\right)^s,$$

则根据 (8.131)，对量 ε_s 有

$$\varepsilon_s \simeq \mu^s \left(\frac{\partial f(t, x)}{\partial x}\right)^{-s} \frac{d^s x}{dt^s}, \tag{8.133}$$

即在边界层外，误差与 μ^s 成正比．考虑到原始刚性方程的解的导数的量值与 μ^{-s} 相比在边界层外是非常小的．因此当 $t \geqslant t_0 + \tau_{BL}$ 时，可以用 (8.133) 来描述 $x(t)$．

作为这种考虑的一个例子为

例 8.3　考虑刚性方程

$$\frac{dx}{dt} = \alpha(t)(x - \varphi) + \frac{d\varphi}{dt}, \quad x(0) = x_0, \tag{8.134}$$

$$\alpha(t) = -10^4 - 3t^2, \quad \varphi(t) = \cos t, \quad t \in [0, 1].$$

它的解可以表成形式

$$x(t) = \exp(-10^4 t - t^3)(x_0 - 1) + \cos t$$

在边界层外 ($\tau_{BL} \sim 10^{-3}$), 解的第一个加项实际上可看成为零, 解 $x(t)$ 接近于 $\cos t$. 下面我们对不同的值 s, 根据公式 (8.129) 和 (8.131) 给出 $\tilde{x}(t)$ 和 ε_s 的值:

$$s = 1,\ \tilde{x}(t) = \cos t - \frac{\sin t}{10^4 + 3t^2},\ |\varepsilon_1| \leqslant c_1 \approx 10^{-4};$$

$$s = 2,\ \tilde{x}(t) = \cos t + \frac{\cos t}{(10^4 + 3t^2)^2 - 6t},$$
$$|\varepsilon_2| \leqslant c_2 \simeq 10^{-8};$$

$$s = 3,\ \tilde{x}(t) = \cos t + \frac{\sin t}{(10^4 + 3t^2)^3 - 18t(10^4 + 3t^2) + 6},$$
$$|\varepsilon_3| \leqslant c_3 \simeq 10^{-12}.$$

这里应该指出,对于应用导数拟稳定性原则, 条件 (8.132) 并不是必要的,而是为了得到简化公式(8.133). 量 $\partial f(t, x)/\partial x$ 在解上是可以变化很大的, 但是由 (8.131) 得到的值 $|\varepsilon_s|$ 仍是小的.

例如对于参数为

$$\alpha(t) = -10^3 - 10^4 t^2$$

或者

$$\alpha(t) = -2 \times 10^4 - 10^4 \sin 10t$$

的方程 (8.134) 的解, 虽然在区间 [0, 1] 上 $\alpha(t)$ 的变化非常大, 但对于任意的 s, 近似 (8.129) 将给出满意的结果.

检查所取的近似的正确性, 有许多不同的方法. 例如可以直接在边界层外积分, (8.83)的解分量应该以某种精度满足(8.126), 也可以对 s 和 $s+1$ 均进行计算来比较它们得到的结果.

若在建立代数关系式 (8.126) 时很难得到高阶导数的明显的表示式, 或者微分方程 (8.83) 的右边不是由解析式表示的, 可以利用 Runge-Kutta 方法所实现的思想来近似这些导数, 且构造 (8.126) 中的函数 $\phi(t, x_1, x_2)$ 的近似 $F(t, h_D, z_1, z_2)$, 而为了降低 (8.83) 的阶, 求解方程

$$F(t,\ h_D,\ z_1,\ z_2)=0. \tag{8.135}$$

当然，为了按照给定的精度确定 $F(t,\ h_D,\ z_1,\ z_2)$ 的值，需要计算若干次 (8.83) 的右函数. 量 z_1 和 z_2 分别是子向量 x_1 和 x_2 的近似. 在 (8.135) 中求解 z_1 时可以应用 Newton 法. 得到了关于 z_2 的微分方程组后，就可以应用通常的数值方法来进行积分.

例如，按照上述处理来代换 $s=2$ 的代数方程式 (8.126) 时，我们可以构造它的近似为

$$\begin{aligned}
&f_1(t+h_D,\ z_1+h_D f_1(t,\ z_1,\ z_2),\ z_2\\
&\quad +h_D f_2(t,\ z_1,\ z_2))-f_1(t,\ z_1,\ z_2)\\
&=h_D\left(\frac{df_1(t,\ z_1,\ z_2)}{dt}+O(h_D)\right)\\
&=0.
\end{aligned} \tag{8.136}$$

一般来说，h_D 与积分步长 h 是无关的，它由 (8.136) 所要满足的给定的精度来确定. 当量 h_D 固定时，与 Runge-Kutta 公式类似，增加计算原始方程组的右函数将能提高精度. 给定 $z_2(t)$，由 (8.136) 用 Newton 法可计算 $z_1(t)$，代入 (8.125)，得到方程组

$$\frac{dz_2}{dt}=f_2(t,\ z_1,\ z_2). \tag{8.137}$$

对于 (8.137) 进行数值积分，例如用显式 Runge-Kutta 方法就可以得到 $z_2(t+h)$. 在计算过程中相应地得到 z_1 的值. Newton 方法的初值可用上一步得到的 z_1 或者用前面几个点的 z_1 值进行外插得到. 在求解过程中 Jacobi 矩阵 $\partial F/\partial z_1$ 变化不大，可以积分若干步后计算一次.

若刚性组 (8.83) 可以分出小参数 μ，写成奇异摄动的形式

$$\mu\frac{du}{dt}=f_1(t,\ u,\ v),\ \ u\in R_u^k, \tag{8.138}$$

$$\frac{dv}{dt}=f_2(t,\ u,\ v),\ \ v\in R_v^{m-k}, \tag{8.139}$$

则可用下面的退化组

$$f_1(t,\ \tilde{u},\ \tilde{v})=0, \tag{8.140}$$

$$\frac{d\tilde{v}}{dt}=f_2(t,\tilde{u},\tilde{v}) \tag{8.141}$$

来得到(8.138)和(8.139)的近似解. Тихонов 给这种处理给出了理论基础. 但是这种处理具有一系列的限制.

首先,明显地分出小参数,并将变量分成"快变的"(u)和"慢变的"(v)并不是对所有的刚性组都是适合的. 另外,即使可以分出导数前的小参数,也可能给不出关于快变变量的信息. 例如由(8.121)和(8.122)给出的方程,两个变量是完全平等的. 按其中任何一个得到的近似(8.140)和(8.141)都不能导出所需要的结果.

第二,虽然微分方程组(8.140)和(8.141)可能是由奇异摄动组通过近似得到的,但可能仍是刚性组. 这是因为独立的快速衰减的特解的个数可能与分出的导数前具有小参数的方程的个数不符合.

第三,由求解(8.140)和(8.141)得到的(8.138)和(8.139)的近似解的精度依赖于量 μ,这个精度可能不是充分小的. 这时在边界层外想寻找按 μ 的幂级数展开的形式解

$$\begin{aligned}u(t,\mu)&=\bar{u}_0(t)+\mu\bar{u}_1(t)+\cdots+\mu^k\bar{u}_k(t)+\cdots,\\ v(t,\mu)&=\bar{v}_0(t)+\mu\bar{v}_1(t)+\cdots+\mu^k\bar{v}_k(t)+\cdots,\end{aligned} \tag{8.142}$$

将(8.142)代入(8.138)和(8.139)中,并将这方程的两边也按 μ 的幂级数展开,且令两边的 μ 的同次幂的系数相等,得到确定 $\bar{u}_i(t)$ 和 $\bar{v}_i(t)$ 的方程. 比较零次幂的系数,我们有

$$f_1(t,\bar{u}_0,\bar{v}_0)=0,$$

$$\frac{d\bar{v}_0}{dt}=f_2(t,\bar{u}_0,\bar{v}_0).$$

因此 $\bar{u}_0$ 和 $\bar{v}_0$ 的上述方程与(8.140)、(8.141)重合. 且我们得到

$$\frac{d\bar{u}_0}{dt}=\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial u}\bar{u}_1+\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial v}\bar{v}_1, \tag{8.143}$$

$$\frac{d\bar{v}_1}{dt}=\frac{\partial f_2(t,\bar{u}_0,\bar{v}_0)}{\partial u}\bar{u}_1+\frac{\partial f_2(t,\bar{u}_0,\bar{v}_0)}{\partial v}\bar{v}_1, \tag{8.144}$$

对于其余的 $\bar{u}_i$ 和 $\bar{v}_i$ 也可以建立类似的关系式.

实际上寻找形式为(8.142)的解是非常困难的,这不仅是对于每一对 $\bar{v}_i$, $\bar{u}_i$ 要求求解 (8.143) 和 (8.144) 型的微分代数方程,而且必须确定初始条件 $\bar{v}_i(0)$. 在 § 1 得到了寻找这些初值的公式,其中要求考虑描述边界层的 Π 函数.

在 § 2 中的方法考虑了对刚性系统 $s=1$ 的拟稳定性近似,但是它的处理依赖于分出小参数,上面所述的困难仍保持.

但是根据导数拟稳定性原则得到的近似 (8.126) 和 (8.127) 没有这些困难,并且不需要将刚性组写成奇异摄动的形式 (8.138) 和 (8.139).

现在我们详细考虑将近似 (8.126) 和 (8.127) 应用到 (8.138) 和 (8.139) 的刚性组的情形. 对方程 (8.138) 微分,得

$$\mu\frac{d^2u}{dt^2}=\frac{\partial f_1(t,u,v)}{\partial u}\frac{f_1(t,u,v)}{\mu}+\frac{\partial f_1(t,u,v)}{\partial v}f_2(t,u,v)+\frac{\partial f_1(t,u,v)}{\partial t},\quad(8.145)$$

两边乘 μ 得

$$\mu^2\frac{d^2u}{dt^2}=\frac{\partial f_1(t,u,v)}{\partial u}f_1(t,u,v)+\mu\left(\frac{\partial f_1(t,u,v)}{\partial v}f_2(t,u,v)+\frac{\partial f_1(t,u,v)}{\partial t}\right).$$

类似地,对 (8.135) 微分 $s-1$ 次,得到

$$\mu^s\frac{d^su}{dt^s}=U_0(t,u,v)+\mu U_1(t,u,v)+\cdots+\mu^{s-1}U_{s-1}(t,u,v),\quad(8.146)$$

将 (8.142) 代入 (8.146) 和 (8.149), 表达式 (8.146) 变成

$$\mu^s\frac{d^s\bar{u}_0}{dt^s}+\mu^{s+1}\frac{d^s\bar{u}_1}{dt^s}+\cdots=F_0(t,\bar{u}_0,\bar{v}_0)+\mu F_1(t,\bar{u}_0,\bar{v}_0,\bar{u}_1,\bar{v}_1)+\cdots,\quad(8.147)$$

令 (8.146) 和 (8.139) 中 μ 的同次幂相等, 得到类似于 (8.143) 和 (8.144) 的 $\bar{u}_i$ 和 $\bar{v}_i$ 的方程. 由于 (8.147) 的左边部分的展开式从

μ^s 开始，因而对于 $i < s$，$\bar{u}_i$ 和 $\bar{v}_i$ 的方程不依赖于 (8.147) 的左边部分的项.

根据导数拟稳定性原则，略去方程 (8.146) 中的第 s 阶导数，得到解的近似式. 它按 μ 的展开式与初始组 (8.138)—(8.139) 的展开式 (8.142) 到 μ^{s-1} 项均是符合的. 为了说明这个事实，考虑 $s=2$ 时方程 (8.138)—(8.139) 的型为 (8.126)—(8.127) 的近似.

先将 $\bar{u}_1$ 和 $\bar{v}_1$ 的方程 (8.143) 进行改写，用 $\bar{u}_0$ 和 $\bar{v}_0$ 的表达式代替导数 $d\bar{u}_0/dt$. 对 $\bar{u}_0$ 和 $\bar{v}_0$ 的代数方程微分，得到

$$\begin{aligned}\frac{df_1(t,\bar{u}_0,\bar{v}_0)}{dt} = &\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial u}\frac{d\bar{u}_0}{dt} \\ &+ \frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial v} f_2(t,\bar{u}_0,\bar{v}_0) \\ &+ \frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial t} \\ = &\,0.\end{aligned}$$

假定矩阵 $\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial u}$ 是非奇异的，有

$$\begin{aligned}\frac{d\bar{u}_0}{dt} = -\left(\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial u}\right)^{-1}\Bigg(&\frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial v} f_2(t,\bar{u}_0,\bar{v}_0) \\ &+ \frac{\partial f_1(t,\bar{u}_0,\bar{v}_0)}{\partial t}\Bigg).\end{aligned}$$

并且方程 (8.143) 变换成

$$\frac{\partial f_1}{\partial u}\bar{u}_1 + \frac{\partial f_1}{\partial v}\bar{v}_1 + \left(\frac{\partial f_1}{\partial u}\right)^{-1}\left(\frac{\partial f_1}{\partial v} f_2 + \frac{\partial f_1}{\partial t}\right) = 0. \quad (8.148)$$

另一方面，令对 u 的二阶导数为零，由 (8.145)，我们有

$$\begin{aligned}\frac{\partial f_1(t,u,v)}{\partial u} f_1(t,u,v) + \mu\Bigg(&\frac{\partial f_1(t,u,v)}{\partial v} f_2(t,u,v) \\ &+ \frac{\partial f_1(t,u,v)}{\partial t}\Bigg) = 0, \quad (8.149)\end{aligned}$$

$$\frac{dv}{dt} = f_2(t, u, v). \tag{8.150}$$

将展开式 (8.142) 代入 (8.149) 和 (8.150)，并让 μ 的零次幂的系数相等，我们得到

$$\frac{\partial f_1(t, \bar{u}_0, \bar{v}_0)}{\partial u} f_1(t, \bar{u}_0, \bar{v}_0) = 0,$$

$$\frac{d\bar{v}_0}{dt} = f_2(t, \bar{u}_0, \bar{v}_0).$$

对第一个方程两边乘上 $\left(\frac{\partial f}{\partial u}\right)^{-1}$ 后得到相应于 (8.138) 和 (8.139) 的 $\bar{u}_0$ 和 $\bar{v}_0$ 的方程组．类似地，令 μ 的一次幂的系数相等，v_1 的微分方程将与 (8.144) 重合，而对于 $\bar{u}_1$ 和 $\bar{v}_1$ 的代数方程组恰好是方程 (8.148) 乘上非奇异矩阵 $\partial f_1/\partial u$.

虽然方程 (8.138) 和 (8.139) 的展开式 (8.142) 的前二项和方程 (8.149)—(8.150) 的展开式的前二项是一致的，但是应该注意根据导数的拟稳定性原则得到近似与按公式 (8.143) 和 (8.144) 的近似 (8.142) 在本质上是不同的．按公式 (8.126) 和 (8.127) 的工作只需对 $\tilde{x}_1$ 求解一个代数方程组 (8.126) 和 $\tilde{x}_2$ 的一个初值．而对于用 (8.143) 和 (8.144) 的情形需要寻找所有的 $\bar{v}_i(0)$ 以及求解若干形式为 (8.143) 的方程．为了寻找边界层外 $\tilde{x}_2$ 的初值，可以通过直接积分完全组 (8.124) 和 (8.125)，如果边界层内方程 (8.124) 和 (8.125) 可线性化，则 $\tilde{x}_2(0)$ 可按公式 (8.112) 计算．

本 章 附 注

§1 的材料主要取自 Васильева 和 Бутузов 的 [114].

§2 是根据 Miranker [88] 编写的.

§3 是根据 Ракитский, устинов 和 Черноруцкий 的 [115] 的第三章编写的.

第九章　隐式 Runge-Kutta 方法

1964 年，Butcher[33] 首先提出了 s 级 $2s$ 阶的隐式 Runge-Kutta 方法．1968 年，Ehle[50] 指出 s 级的 $2s$ 阶的隐式 Runge-Kutta 方法是 A 稳定的．由于这类方法是 A 稳定的，又能达到那么高阶的精确度，所以吸引了许多人从事这方面的研究．为了进一步探讨该方法的稳定性质，1975 年，Butcher[34] 提出了 B 稳定性的概念．1979 年，Burrage 和 Butcher[39] 又推广了这个概念，相应地建立了研究该方法的稳定性的代数判别理论．至于隐式 Runge-Kutta 方法如何在数值计算中实现，又提出了一系列的方法．这方面的内容我们将在下一章中讨论．在这一章中，我们主要介绍隐式 Runge-Kutta 公式，讨论它的 A 稳定性以及其它的数值稳定性质．

§1　隐式 Runge-Kutta 公式

数值积分 m 维初值问题

$$y' = f(t, y), \quad y(t_0) = y_0 \tag{9.1}$$

的 s 级的 Runge-Kutta 公式的一般形式为

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i K_i, \tag{9.2}$$

$$K_i = f\left(t_n + c_i h,\ y_n + h\sum_{j=1}^{s} a_{ij}K_j\right), \tag{9.3}$$

其中 $t_n = t_0 + nh\ (n = 0, 1, 2, \cdots)$ 为时间轴上离散点列，h 为积分步长，y_n 为解 $y(t_n)$ 的近似值，$c_1, c_2, \cdots, c_s$ 称为 Runge-Kutta 方法的节点，$b_1, b_2, \cdots, b_s$ 为权系数，$A = (a_{ij})\ (i, j=$

$1, 2, \cdots, s$) 称为方法的系数矩阵，满足条件 $c_i = \sum_{j=1}^{s} a_{ij}$, ($i = 1, 2, \cdots, s$). 于是，Runge-Kutta 方法 (9.2)、(9.3) 可由如下的系数表来描述：

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & & \cdots\cdots & \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array},$$

使用矩阵和向量的记号，上表可写为

$$\begin{array}{c|c} C & A \\ \hline & B^T \end{array},$$

其中 C、B 表示列向量，上标 T 表示转置. 如果 A 是一个主对角元素均为零的下三角形矩阵，相应的 Runge-Kutta 公式是显式的. 这时，用 (9.3) 式计算 K_i 时，其右端只含 $K_1, K_2, \cdots, K_{i-1}$. 如果 A 是一个主对角元素为非零的下三角形矩阵，相应的 Runge-Kutta 公式称为半隐式的. 这时 (9.3) 式右端含有 $K_1, K_2, \cdots, K_i$, 求 K_i 时要解一个只含 K_i 的方程组. 如果矩阵 A 为一般的 s 阶矩阵，相应的公式称为隐式的. 这时 (9.3) 式的右端一般都含有全部的 K_i, $i = 1, 2, \cdots, s$. 求 K_i 时要解含 $K_1, \cdots K_s$ 的方程组.

Kutta[65] 得到的三级三阶的显式 Runge-Kutta 公式，它对应

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & -1 & 2 & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}.$$

于还可举出三级的半隐式和隐式 Runge-Kutta 公式的例子，它们分别为

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\
0 & 0 & 1 & 0 \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array},
\qquad
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\
1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}.
$$

为了求出具体的 Runge-Kutta 公式,需要确定参数 C、B、A. 通常有两种方法. 一种方法将(9.3)在 (t_n, y_n) 点展开, 代入(9.2)中, 并与 $y(t_n+h)$ 在 t_n 点的 Taylor 展开式相比较, 来确定 C、B 和 A. 这里 $y(t)$ 是微分方程的解. 另一种方法是将微分方程化成等价的积分方程, 用数值积分求得 y_{n+1} 的表达式,把它和(9.3)的展式代入(9.2)中的结果相比较,以确定诸参数.

基于数值积分的 Runge-Kutta 公式有许多的形式, 下面概述[12], [33], [71], [51], [23], [24], [42] 中介绍的三种. 引进 s 阶矩阵的符号: $W=(c_i^{j-1})$, $C=(c_i^j/j)$, $N=(1/i)$, $D=\mathrm{diag}(b_i)$.

1. 基于 Gauss 型求积公式的 G 类方法(见 Butcher[33])

Butcher 指出 $c_1, c_2, \cdots c_s$ 为 $P_s(2C-1)=0$ 的根, 其中 $P_s(t)$ 为 $[0,1]$ 上的 s 次 Legendre 多项式, $0<c_i<1$ $(i=1, 2, \cdots, s)$, 求 s 级的 $2s$ 阶的 Runge-Kutta 公式的参数的步骤如下:

1) 求出 s 次的 Legendre 多项式 $P_s(2C-1)$ 的 s 个零点, $c_1, c_2, \cdots, c_s$.

2) 由线性方程组

$$\sum_{j=1}^{s} b_j c_j^{k-1}=1/k, \quad k=1, 2, \cdots, s$$

来确定系数 b_j $(j=1, 2, \cdots, s)$.

3) 计算系数矩阵 A

$$A=CW^{-1}.$$

在这个基础上，Butcher[33] 提出一系列的当 $s=1,2,\cdots,5$ 和 $p=2s$ 时的隐式 Runge-Kutta 的 Gauss 型公式。下面我们仅给出几个常用的 Runge-Kutta 公式的系数表[见[71]]。

$s=1$，$p=2$ 的隐式 Runge-Kutta 的 Gauss 型公式

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}. \tag{9.4}$$

$s=2$，$p=4$ 的隐式 Runge-Kutta 的 Gauss 型公式

$$\begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array}. \tag{9.5}$$

$s=3$，$p=6$ 的隐式 Runge-Kutta 的 Gauss 型公式

$$\begin{array}{c|ccc} (5-\sqrt{15})/10 & 5/36 & (10-3\sqrt{15})/45 & (25-6\sqrt{15})/180 \\ 1/2 & (10+3\sqrt{15})/72 & 2/9 & (10-3\sqrt{15})/72 \\ (5+\sqrt{15})/10 & (25+6\sqrt{15})/180 & (10+3\sqrt{15})/45 & 5/36 \\ \hline & 5/18 & 4/9 & 5/18 \end{array}. \tag{9.6}$$

2. 基于 Radau 求积公式的 Runge-Kutta 方法

Ehle[51] 和 Axelsson[23] 研究了这种类型的方法。这类方法求参数 C、B 和 A 的步骤如下：

1）求多项式 $P_s(t)-P_{s-1}(t)$ 的零点 $c_1,c_2,\cdots,c_s$。并指定 $c_1>0$，$c_s=1$。

2）计算系数 b_k

$$b_k=\int_0^1 \frac{P_s(t)-P_{s-1}(t)}{(t-c_k)[P_s'(c_k)-P_{s-1}'(c_k)]}dt,$$

$$k=1,2,\cdots,s.$$

3）计算系数矩阵 A

$$A=CW^{-1}.$$

下面我们给出两个隐式 Runge-Kutta 的 Radau 公式的系数表（见[71]）：

$s=2$，$p=3$ 的隐式 Runge-Kutta 的 Radau 公式，即半隐式

的

$$
\begin{array}{c|cc}
\frac{1}{3} & \frac{1}{3} & 0 \\
1 & 1 & 0 \\
\hline
 & \frac{3}{4} & \frac{1}{4}
\end{array}
\tag{9.7}
$$

$s=3$, $p=5$ 的隐式 Runge-Kutta 的 Radau 公式

$$
\begin{array}{c|ccc}
(4-\sqrt{6})/10 & (24-\sqrt{6})/120 & (24-11\sqrt{6})/120 & 0 \\
(4+\sqrt{6})/10 & (24+11\sqrt{6})/120 & (24+\sqrt{6})/120 & 0 \\
1 & (6-\sqrt{6})/12 & (6+\sqrt{6})/12 & 0 \\
\hline
 & (16-\sqrt{6})/36 & (16+\sqrt{6})/36 & 1/9
\end{array}
\tag{9.8}
$$

3. 基于 Lobatto 求积公式的 Runge-Kutta 方法

Ehle[51]、Axelsson[23],[24], Chipman[42] 研究了这种类型的方法. 这类方法求参数 C、B 和 A 的步骤如下:

1) 求多项式 $P_s(t)-P_{s-2}(t)$ 的零点 $c_1, c_2, \cdots, c_s$, 并指定 $c_1=0$, $c_s=1$.

2) 计算权系数 b_k

$$
b_k=\int_0^1 \frac{P_s(t)-P_{s-2}(t)}{(t-c_k)[P_s'(c_k)-P_{s-2}'(c_k)]}dt,
$$

$$
k=1, 2, \cdots, s.
$$

3) 计算系数矩阵 $A=D^{-1}(W^T)^{-1}(N-C)^TD$.

下面我们由 [71] 给出几个隐式 Runge-Kutta 的 Lobatto 公式的系数表:

$s=2$, $p=2$ 的隐式 Runge-Kutta 的 Lobatto 公式, 即显式

$$
\begin{array}{c|cc}
0 & 0 & 0 \\
1 & 1 & 0 \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}.
\tag{9.9}
$$

$s=3$, $p=4$ 的隐式 Runge-Kutta 的 Lobatto 公式, 此即半隐式公式

$$
\begin{array}{c|ccc}
0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\
1 & 0 & 1 & 0 \\
\hline
 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6}
\end{array}. \tag{9.10}
$$

$s=4$, $p=6$ 的隐式 Runge-Kutta 的 Lobatto 公式

$$
\begin{array}{c|cccc}
0 & 0 & 0 & 0 & 0 \\
(5-\sqrt{5})/10 & (5+\sqrt{5})/60 & 1/6 & (15-7\sqrt{5})/60 & 0 \\
(5+\sqrt{5})/10 & (5-\sqrt{5})/60 & (15+7\sqrt{5})/60 & 1/6 & 0 \\
1 & 1/6 & (5-\sqrt{5})/12 & (5+\sqrt{5})/12 & 0 \\
\hline
 & 1/12 & 5/12 & 5/12 & 1/12
\end{array}. \tag{9.11}
$$

$s=5$, $p=8$ 的隐式 Runge-Kutta 的 Lobatto 公式

$$
\begin{array}{c|ccccc}
0 & 0 & 0 & 0 & 0 & 0 \\
(7-\sqrt{21})/14 & 1/14 & 1/9 & (13-3\sqrt{21})/63 & (14-3\sqrt{21})/126 & 0 \\
1/2 & 1/32 & (91+21\sqrt{21})/576 & 11/72 & (91-21\sqrt{21})/576 & 0 \\
(7+\sqrt{21})/14 & 1/14 & (14+3\sqrt{21})/126 & (13+3\sqrt{21})/63 & 1/9 & 0 \\
1 & 0 & 7/18 & 2/9 & 7/18 & 0 \\
\hline
 & 1/20 & 49/180 & 16/45 & 49/180 & 1/20
\end{array}. \tag{9.12}
$$

现在我们来证明 s 级的隐式 Runge-Kutta 公式具有 $2s-v$ 阶的精确度，其中整数 $v=0, 1, 2$. 这里我们不采用 1964 年，Butcher[33] 中的证法，而采用 1969 年，Axelsson[23] 中的证法，因为这种证明方法更加初等一些.

对于给定的初值问题

$$
y' = f(t, y),\ y(0) = \eta,\ 0 \leqslant t \leqslant T, \tag{9.13}
$$

其中 y、f 均为 m 维向量. Axelsson 研究了如下形式的数值解公式

$$
y_{i,n+1} = y_{s,n} + h\sum_{k=1}^{s} a_{ik} f((n+u_k)h, y_{k,n+1}), \tag{9.14}
$$

$$
i=1, 2, \cdots, s;\ n=0, 1, 2, \cdots,
$$

其中 $y_{s,0}=\eta$, $h>0$ 是积分步长, s 为一个积分步内所取的节点个数，一般为常数. 节点 u_k 和求积系数 a_{ik} 按照如下的方式

确定.

令 Π_m 为所有的次数小于等于m的多项式集合. 令P_m是次数等于m的在区间[0, 1]上正交的 Legendre 多项式, 归格化使得 $P_m(1)=1$, 且具有最高的正系数. 又令

$$Q_s(t)=P_s(t)+aP_{s-1}(t)+bP_{s-2}(t), \tag{9.15}$$

它是 Legendre 多项式的一个线性组合, 其中 a、b 为线性组合的实系数. Shohat[23] 指出, 若 $b\leqslant 0$, 则它有 s 个互异的实零点,

$$u_1<u_2<\cdots<u_s.$$

在本章的前头, 节点符号取为 c_i, $i=1,2,\cdots,s$. 这里为了下面定理叙述方便, 把节点符号改为用 u_i 来表示, $i=1,2,\cdots,s$. 在以下的讨论中, 我们假定 $u_1\geqslant 0$, $u_s=1$. 令

$$L_k(t)=\frac{Q_s(t)}{(t-t_k)Q'(t_k)},\quad k=1,2,\cdots,s.$$

又令求积系数

$$a_{ik}=\int_0^{u_i}L_k(x)dx,\quad i,k=1,2,\cdots,s, \tag{9.15$_1$}$$

和截断误差

$$R_i(f)=\int_0^{u_i}f(x)\,dx-\sum_{k=1}^{s}a_{ik}f(u_k).$$

那么因为这是一个内插求积公式, 所以

$$R_i(f)=0,\quad f\in\Pi_{s-1}.$$

因此, 系数

$$a_k=a_{sk}=\int_0^1 L_k(x)dx \tag{9.16}$$

与经典的求积系数相等. Axelsson 指出, 若 $b\leqslant 0$, 则它们都是正的, 且

$$R_s(f)=0,\quad f\in\Pi_{2s-v-1},$$

其中

$$v=\begin{cases}2 & \text{如果 } b\neq 0,\\ 1 & \text{如果 } a\neq 0,\ b=0,\\ 0 & \text{如果 } a=b=0,\end{cases}$$

(后面这种情况不适合于 $u_s=1$ 的情形).

我们看到,在公式 (9.14) 中,令 $t_n=nh$,

$$k_i=f\left(t_n+u_ih,\ y_{s,n}+h\sum_{j=1}^{s}a_{ij}k_j\right),$$

在公式 (9.16) 中,又令 $b_i=a_k=a_{sk}$, $i=k=1,2,\cdots,s$, 并且把公式 (9.14) 中 $y_{s,n}$, $y_{s,n+1}$ 的下标 s 省掉,把 u_i 换成 c_i, 则公式 (9.14) 就有形式

$$y_{n+1}=y_n+h\sum_{i=1}^{s}b_ik_i, \tag{9.17}$$

其中

$$k_i=f\left(t_n+c_ih,\ y_n+h\sum_{j=1}^{s}a_{ij}k_j\right).$$

这就是通常的 s 级隐式 Runge-Kutta 方法的形式 (9.2) 和 (9.3). 特别,若 u_i 取为多项式 $Q_s(t)=P_s(t)-P_{s-1}(t)$ 的零点,且 $0<u_1<u_2<\cdots<u_s=1$,则上述方法对应于隐式 Runge-Kutta 方法的 Radau 求积公式. 若 u_i 取为多项式 $Q_s(t)=P_s(t)-P_{s-2}(t)$ 的零点,且 $0=u_1<u_2<\cdots<u_s=1$,则上述方法对应于隐式 Runge-Kutta 方法的 Lobatto 求积公式(见 [23]、[24]).

下面我们来讨论方法的误差阶. 令

$$J(t)=J(t,f,y)=\partial f/\partial y(t,y),$$

设 I 是 $s\times m$ 阶的单位矩阵,并令

$$\hat{J}=\hat{J}(h)=\begin{pmatrix}a_{11}J(hu_k) & \cdots & a_{1s}J(hu_k)\\ & \cdots\cdots\cdots & \\ a_{s1}J(hu_k) & \cdots & a_{ss}J(hu_k)\end{pmatrix}=A\otimes J(hu_k).$$

对于隐式 Runge-Kutta 公式的精确度,我们有:

定理 9.1 如果对于 $h\leqslant T$, $I-h\hat{J}$ 是非奇异的, $\|\partial^2 f/\partial y^i\partial y^j\|$, $1\leqslant i,j\leqslant m$, 在充分大的区域内是有界的,且 $(\partial f^i/\partial y^j)(t,y)\in C^{s-v-1}[0,T)$ 和 $y(t)\in C^{2s-v+1}[0,T)$, 那么

$$\|y(h)-y_s\|=O(h^{2s-v+1}),\quad h\leqslant h_0.$$

证明. 不失一般性,我们考虑自守系统

$$y' = f(y), \ y(0) = \eta, \ 0 \leqslant t < T.$$

在所考虑的区域内,令

$$\|\partial^2 f(y)/\partial y^i \partial y^j\| \leqslant K.$$

假定

$$e_i = y(hu_i) - y_i,$$

且令块向量 $e = [e_i] = [e_1, \cdots, e_s]^T$.

这样

$$\begin{aligned} e_i &= h\int_0^{u_i} f(y(hu))du - h\sum_{k=1}^{s} a_{ik} f(y(hu_k)) \\ &\quad + h\sum_{k=1}^{s} a_{ik}[f(y(hu_k)) - f(y_k)] \\ &= h\sum_{k=1}^{s} a_{ik}[J(hu_k)e_k + O(\|e_k\|^2)] + R_i(z), \end{aligned}$$

其中

$$R_i(g) = \int_0^{u_i} g(u)du - \sum_{k=1}^{s} a_{ik} g(u_k) \tag{9.18}$$

和

$$z(u) = hy'(hu).$$

于是有

$$e = h\hat{J}e + R(z) + h\max_k O(\|e_k\|^2),$$

其中 $R(z)$ 是块向量 $[R_i(z)]$. 因为 R_i 是线性算子, 由 (9.15_1) 知 $R_i = (u^q) = 0$ $(q \leqslant s-1)$, 因而

$$R_i(g) = R_i\left(\frac{u^s}{s!} g^{(s)}(\xi(u))\right), \ 0 < \xi(u) < u,$$

又由题设 $(I - h\hat{J})^{-1}$ 存在, 所以

$$\|e_i\| = O(R_i(z)) = O(h^{s+1}).$$

根据上面的结果,我们得到

$$\begin{aligned} e &= h\hat{J}e + R(z) + O(h^{2s+3}) \\ &= (h\hat{J})^2 e + h\hat{J}R(z) + R(z) + O(h^{2s+3}), \end{aligned}$$

将上式左端的 e 逐次代入右端并通过归纳，得

$$e=(h\hat{f})^{p}e+\sum_{l=0}^{p-1}(h\hat{f})^{l}R(z)+O(h^{2s+3}). \qquad (9.19)$$

为完成余下的证明，需要下面两个引理.

现在在 (9.15) 中令

$$\begin{aligned}Q_s(u)&=P_s(u)+aP_{s-1}(u)+bP_{s-2}(u)\\&=\alpha_0(u^s+\alpha_1u^{s-1}+\cdots+\alpha_s)\ (\alpha_0>0),\end{aligned} \qquad (9.19_1)$$

并令

$$\omega_s(u)=\frac{1}{\alpha_0}Q_s(u). \qquad (9.19_2)$$

引理 9.1 假定 $R_i(g)$ 由 (9.18) 式定义，那么

$$\begin{aligned}R_i(u^{s+q})&=\int_0^{u_i}s^{*q}\omega_s(s^*)ds^*\\&\quad-\sum_{j=1}^{q}\alpha_jR_i(u^{s+q-j}),\ q=0,1,\cdots,s,\end{aligned}$$

且

$$R_s(u^{s+q})=0,\ q\leqslant s-v-1.$$

证明 因为 R_i 是线性算子，使

$$R_i(u^q)=0,\ q=0,1,\cdots,s-1.$$

由 (9.19_2) 知

$$\omega_s(u_k)=0,\ k=1,2,\cdots,s.$$

由 (9.18) 得到

$$R_i(u^s)=R_i(\omega_s)=\int_0^{u_i}\omega_s(s^*)ds^*,$$

又由 (9.19_1) 和 (9.19_2) 知

$$\begin{aligned}&R_i[u^q(u^s+\alpha_1u^{s-1}+\cdots+\alpha_qu^{s-q})]\\&\quad+R_i(\alpha_{q+1}u^{s-1}+\cdots+\alpha_su^q)\\&=R_i(u^q\omega_s(u))\\&=\int_0^{u_i}u^q\omega_s(u)\,du+\sum_{k=1}^{s}a_{ik}u_k^q\omega_s(u_k),\end{aligned}$$

由此可得

$$R_i(u^{s+q})+\sum_{j=1}^{q}\alpha_j R_i(u^{s+q-j})=\int_0^{u_i}s^{*q}\omega_s(s^*)ds^*.$$

引理 9.2 如果 $a_k=\int_0^1 L_k(s^*)ds^*$，那么对于 $f_q\in\Pi_q$，当 $q+r\leqslant 2s-v-2$ 时，有

$$\sum_{k=1}^{s}a_k(1-u_k)^r\int_0^{u_k}f_q(s^*)ds^*=\int_0^1\frac{(1-u)^{r+1}}{r+1}f_q(u)du.$$

特别，对于 $a_{ik}=\int_0^{u_i}L_k(s^*)ds^*$ 有

$$\sum_{k=1}^{s}a_k a_{kj}(1-u_k)^r=a_j\frac{(1-u_j)^{r+1}}{r+1},\quad r\leqslant s-v-1.$$

证明 这里 $L_k(t)=\dfrac{Q_s(t)}{(t-t_k)Q_s'(t)}$，由 (9.16) 知 $a_k=a_{sk}$。若令 $\varphi(u)=(1-u)^r\int_0^u f_q(v)dv$，它是 $q+r$ 次多项式，由引理 9.1，因 $r+(q+1)\leqslant 2s-v-1$，故有

$$\begin{aligned}\sum_{k=1}^{s}a_k(1-u_k)^r\int_0^{u_k}f_q(s^*)ds^*&=\sum_{k=1}^{s}a_{sk}\varphi(u_k)ds^*\\&=\int_0^1\varphi(s^*)ds^*,\end{aligned}$$

用分部积分得本引理的第一等式。在本引理的第一等式中取 $f_q(s^*)=L_k(s^*)$，则得

$$\sum_{k=1}^{s}a_k a_{kj}(1-u_k)^r=\int_0^1\frac{(1-u)^{r+1}}{r+1}L_j(u)du.$$

当 $(r+1)+(s-1)\leqslant 2s-v-1$ 时，由引理 9.1，得

$$\int_0^1\frac{(1-u)^{r+1}}{r+1}L_j(u)du=a_j\frac{(1-u_j)^{r+1}}{r+1},$$

引理 9.2 证毕。

现在我们再继续证明定理 9.1。由 Taylor 展开式，取 $hu=h-(1-u)h$，有

$$J(hu) = \sum_{r=0}^{s-v-2} (-1)^r h^r \frac{(1-u)^r}{r!} J^r(h) + O(h^{s-v-1}). \quad (9.20)$$

考虑 (9.19) 中的块向量 e 的第 s 个分量，将 $y'(hu)$ 展开，由引理 9.1，得

$$R_s(z) = hR_s(y'(hu))$$
$$= h^{2s-v+1} R_s\left(\frac{u^{2s-v}}{(2s-v)!} y^{(2s-v+1)}(h\theta(u))\right),$$
$$0 < \theta(u) < u,$$

因而，

$$R_s(z) = O(h^{2s-v+1}).$$

另外，根据引理 9.1 和 9.2，

$$h \sum_{k=1}^{s} a_k h^r \frac{(1-u_k)^r}{r!} R_k((hu)^{s+q})$$
$$= h^{s+q+r+1} \sum_{k=1}^{s} a_k \frac{(1-u_k)^r}{r!}$$
$$\cdot \left\{\int_0^{u_k} s^{*q} \omega_s(s^*) ds^* - \sum_{j=1}^{q} \alpha_j R_k(u^{s+q-j})\right\}$$
$$= h^{s+q+r+1} \int_0^1 \left\{\frac{(1-u)^r}{r!} \int_0^u (s^{*q}\right.$$
$$\left. + (\text{低次项}))\omega_s(s^*)ds^*\right\} du$$
$$= h^{s+q+r+1} \int_0^1 \left\{\frac{(1-u)^{r+1}}{r+1} (u^q\right.$$
$$\left. + (\text{低次项}))\right\}\omega_s(u)du$$
$$= 0, \quad 0 \leqslant q + r \leqslant s - v - 2,$$

上式等于零是根据正交性

$$\int_0^1 \omega_s(u) f(u) du = 0, \quad f \in \Pi_{s-v-2}$$

得出来的. 于是，由 (9.20)

$$h\sum_{k=1}^{s} a_k J(hu_k) R_k((hu)^{s+q}) = O(h^{2s-v}),$$

$$0 \leqslant q \leqslant s-v-2,$$

并由

$$z(u) = hy'(hu) - h\sum_{r=0}^{2s-v-2} \frac{(hu)^r}{r!} y^{(r+1)}(0) + O(^{2s-v}),$$

我们得到

$$R(z) = h\sum_{r=s}^{2s-v-2} \frac{y^{(r+1)}(0)}{r!} R((hu)^r) + O(h^{2s-v}).$$

所以，块向量 $h\hat{J}R(z)$ 的第 s 个分量是 $O(h^{2s-v+1})$.

现在考虑(9.19)的下一项(即 $l=2$). 由引理 9.2 我们得到

$$\begin{aligned}
&h^2\sum_{k,j=1}^{s} a_k a_{kj} J(hv_k) J(hu_j) \\
&= h^2\sum_{j=1}^{s}\left\{\sum_{r=0}^{s-v-3}(-1)^r\frac{h^r}{r!}J^{(r)}(h)\right. \\
&\quad\left.\cdot\sum_{k=1}^{s} a_k a_{kj}(1-u_k)^r + O(h^{s-v-2})\right\} J(hu_j) \\
&= \sum_{r=0}^{s-v-3}(-1)^r J^{(r)}(h) h\sum_{j=1}^{s} a_j h^{r+1} \\
&\quad\cdot\frac{(1-u_j)^{r+1}}{(r+1)!} J(hu_j) + O(h^{s-v}).
\end{aligned}$$

同理我们得到

$$h^2\sum_{k,j=1}^{s} a_k a_{kj} J(hu_k) J(hu_j) R_j((hu)^{s+q}) = O(h^{2s-v}),$$

$$0 \leqslant q \leqslant s-v-3.$$

因此，块向量 $(h\hat{J})^2R(z)$ 的第 s 个分量也是 $O(h^{2s-v+1})$. 按照同样的方法，我们得到 $(h\hat{J})^lR(z)$ 第 s 个分量是 $O(h^{2s-v+1})$, 如果 $s-v-l-1\geqslant 0$, 即 $l\leqslant s-v-1$.

于是，最后得到

$$e_s = O(h^{2s-v+1}).$$

这是一个积分步后的误差，即局部截断误差．由 Henrici[62] 的文章可得到在 $[0, T)$ 内的整体误差是 $O(h^{2s-\nu})$.

§2 隐式 Runge-Kutta 方法的 A 稳定性

对非线性方程 (9.1) 直接进行其数值解的稳定性分析一般是比较困难的．1963 年，Dahlquist[48] 针对试验方程 $y' = \lambda y$，提出了常微分方程初值问题数值积分方法的 A 稳定性概念，其中 λ 是复数，$\mathrm{Re}(\lambda) < 0$．通常这是作为分析数值积分方法稳定性的基础．我们把隐式 Runge-Kutta 公式 (9.2) 和 (9.3) 应用于上述试验方程，并记

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_s \end{pmatrix}, \quad K = \begin{pmatrix} K_1 \\ \vdots \\ K_s \end{pmatrix}, \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

得到

$$K = \lambda e y_n + \lambda h A K, \tag{9.21}$$

即

$$(I - \lambda h A)K = \lambda e y_n.$$

如果 $(I - \lambda h A)$ 是非奇异的矩阵，则有

$$K = \lambda(I - \lambda h A)^{-1} e y_n,$$

所以

$$y_{n+1} = y_n + h b^T K = (1 + \lambda h b^T (I - \lambda h A)^{-1} e) y_n.$$

令 $z = h\lambda$，$R(z) = 1 + z b^T (I - zA)^{-1} e$，并称 $R(z)$ 为 Runge-Kutta 方法的传播函数（或特征函数）．由数值稳定性定义，对于 $h > 0$，若 $\mathrm{Re}(z) < 0$，有 $|R(z)| < 1$，则方法是 A 稳定的．1968 年，Ehle[50] 第一次指出 s 级的 $2s$ 阶的隐式 Runge-Kutta 方法是 A 稳定的．例如 2 级 4 阶的隐式 Runge-Kutta 的 Gauss 型公式，即公式 (9.5) 有

$$a_{11} = \frac{1}{4}, \qquad a_{12} = \frac{1}{4} - \sqrt{3}/6,$$

$$a_{21} = \frac{1}{4} + \sqrt{3}/6, \qquad a_{22} = \frac{1}{4},$$

$$b_1 = \frac{1}{2}, \qquad b_2 = \frac{1}{2}.$$

因此，对于 $y' = \lambda y$，有

$$K_1 = h\lambda\left(y_n + \frac{K_1}{4} + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)K_2\right),$$

$$K_2 = h\lambda\left(y_n + \left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)K_1 + \frac{K_2}{4}\right),$$

利用 Cramer 法则解出 K_1，K_2，得到

$$K_1 = h\lambda y_n\left(1 - \frac{h\lambda\sqrt{3}}{6}\right)\Big/\Delta,$$

$$K_2 = h\lambda y_n\left(1 + \frac{h\lambda\sqrt{3}}{6}\right)\Big/\Delta,$$

其中

$$\Delta = 1 - \frac{1}{2}(h\lambda) + \frac{(h\lambda)^2}{12}.$$

最后，

$$\begin{aligned} y_{n+1} &= y_n + h(b_1K_1 + b_2K_2) \\ &= \left(\frac{1 + \frac{1}{2}h\lambda + \frac{(h\lambda)^2}{12}}{1 - \frac{1}{2}h\lambda + \frac{(h\lambda)^2}{12}}\right)y_n \\ &= R(h\lambda)y_n \\ &= R(z)y_n \quad (z = h\lambda). \end{aligned}$$

显然，当 $\mathrm{Re}(z) < 0$ 时，$|R(z)| < 1$，该方法是 A 稳定的.

下面我们来比较详细地讨论隐式 Runge-Kutta 方法的 A 稳定性.

在第四章，我们介绍了 Padé 近似，引进了指数函数有理近似的可接受性定义，即假设 $R(z)$ 为指数函数 $\exp(z)$ 的有理近似. 1) 若 $\mathrm{Re}(z) < 0$，有 $|R(z)| < 1$，则称 $R(z)$ 为 A 可接

受的；2）若对于所有负实数 z，有 $|R(z)|<1$，则称 $R(z)$ 为 A_0 可接受的；3）若 $R(z)$ 是 A 可接受的，且当 $\mathrm{Re}(z)\to-\infty$ 时，有 $|R(z)|\to 0$，则称其为 L 可接受的．显然，由第一章给出的数值稳定性定义可以知道：

当 Runge-Kutta 方法的传播函数 $R(z)$ 逼近 $\exp(z)$ 为 A 可接受的，则方法是 A 稳定的；为 A_0 可接受的，则方法是 A_0 稳定的；为 L 可接受的，则方法是 L 稳定的．

1973 年，Ehle[52] 证明，隐式 Runge-Kutta 方法的传播函数 $R(z)$ 是指数函数 $\exp(z)$ 的对角 Padé 近似或第一、二下对角 Padé 近似．所以该方法是 A 稳定的或 L 稳定的．也就是说，如果 $R(z)$ 是一个有理函数，分子的次数为 $s-d$，分母的次数为 s，当 $z\to 0$ 时，对于 $d=0,1,2$，有 $R(z)-e^z=O(z^{2s-d+1})$，那么对于所有左半平面的复数 z，即对于 $\mathrm{Re}(z)<0$，有 $|R(z)|<1$。

Ehle 关于这个结果的证明比较长．这里我们叙述 1977 年，Butcher[36] 给出的一个更加简短的证明．

对于给定的 $z\neq 0$，若令 $a_m=(4m-2)/z,(m=1,2,\cdots,s-1)$ 和 $a_s=(4s-2)/z$（如果 $d=0$），$a_s=(4s-2)/z-1$（如果 $d=1$），$a_s=(2s-2)/z-1$（如果 $d=2$），那么我们有

引理 9.3 $R(z)=1+2/(-1+a_1+1/(a_2+1/(a_3+\cdots+1/(a_{s-1}+1/a_s)\cdots)))$

证明 众所周知（例如见 G. A. Baker [24]），连分式 $\alpha_0+\beta_1/(\alpha_1+\beta_2/(\alpha_2+\cdots))$ 各阶渐近分式 $A_m/B_m\ (m=0,1,2,\cdots)$ 为

$$A_0/B_0=\alpha_0/1,\quad A_1/B_1=(\alpha_1\alpha_0+\beta_1)/\alpha_1,$$

$A_2/B_2=(\alpha_2(\alpha_1\alpha_0+\beta_1)+\beta_2\alpha_0)/(\alpha_2\alpha_1+\beta_2),\cdots$．一般，有关系式

$$\left.\begin{aligned}A_m&=\alpha_mA_{m-1}+\beta_mA_{m-2}\\B_m&=\alpha_mB_{m-1}+\beta_mB_{m-2}\end{aligned}\right\}\quad m=2,3,\cdots,$$

由于$m+1$阶的近似分式 A_{m+1}/B_{m+1} 等于把 A_m/B_m 中的α_m换成 $\alpha_m+\beta_{m+1}/\alpha_{m+1}$. 故用数学归纳法可以证明以上的两个等式成立.

以 N_0/D_0, $(z^{-m}N_m)/(z^{-m}D_m)$ $(m=1,2,\cdots,s)$ 表示连分式 $R(z)$ 的各阶渐近分式. 利用 $\alpha_0=1$, $\alpha_1=-1+a_1$, $\alpha_2=a_2$, $\alpha_3=a_3,\cdots$; $\beta_1=2$, $\beta_2=1$, $\beta_3=1,\cdots$; 得到

$$N_0=D_0=1,\ N_1=2+z,\ D_1=2-z,$$

$$\left.\begin{aligned}N_m&=(4m-2)N_{m-1}+z^2N_{m-2}\\D_m&=(4m-2)D_{m-1}+z^2D_{m-2}\end{aligned}\right\}\ m=2,3,\cdots,s-1,$$

$$\left.\begin{aligned}N_s&=(4s-2)N_{s-1}+z^2N_{s-2}\\D_s&=(4s-2)D_{s-1}+z^2D_{s-2}\end{aligned}\right\}\ d=0,$$

$$\left.\begin{aligned}N_s&=(4s-2-z)N_{s-1}+z^2N_{s-2}\\D_s&=(4s-2-z)D_{s-1}+z^2D_{s-2}\end{aligned}\right\}\ d=1,$$

$$\left.\begin{aligned}N_s&=(2s-2-z)N_{s-1}+z^2N_{s-2}\\D_s&=(2s-2-z)D_{s-1}+z^2D_{s-2}\end{aligned}\right\}\ d=2.$$

在 $s=d=1$ 的特殊情形中, $N_1=2$, $D_1=2-2z$.

对于 $m=2,3,\cdots,s-1$ (在 $d=0$ 的情形,也包括 $m=s$) 利用归纳法可以证明

$$N_m=\sum_{j=0}^{m}\frac{(2m-j)!}{j!(m-j)!}z^j, \tag{9.22}$$

$$D_m=\sum_{j=0}^{m}\frac{(2m-j)!}{j!(m-j)!}(-z)^j. \tag{9.23}$$

对于 $d=1$ 的情形,有

$$N_s=2\sum_{j=0}^{s-1}\frac{(2s-1-j)!}{j!(s-1-j)!}z^j, \tag{9.24}$$

$$D_s=2s\sum_{j=0}^{s}\frac{(2s-1-j)!}{j!(s-j)!}(-z)^j. \tag{9.25}$$

当 $d=2$ 时,有

$$N_s=2\sum_{j=0}^{s-2}\frac{(2s-2-j)!}{j!(s-2-j)!}z^j, \tag{9.26}$$

$$D_s = 2s(s-1)\sum_{j=0}^{s}\frac{(2s-2-j)!}{j!(s-j)!}(-z)^j. \qquad (9.27)$$

当 $d=0$ 时、$R(z)=N_s/D_s$ 是通常的对角 Padé 近似. 当 $d=1,2$ 时 $R(z)=\frac{1}{2}N_s\Big/\frac{1}{2}D_s$ 恰好是第一、二下对角 Padé 近似. 引理证毕.

现在利用引理中给出的 $R(z)$ 表达式证明如下定理.

定理 9.2 如果 $\mathrm{Re}(z)<0$, 那么 $|R(z)|<1$.

证明 对于给定的 s 和 $m=1,2,\cdots,s$, 令

$$A_m = a_m + 1/(a_{m+1} + \cdots + 1/(a_{s-1} + 1/a_s)\cdots),$$

因而有

$$A_s = a_s,\ A_m = a_m + 1/A_{m+1}\ (m=1,2,\cdots,s-1).$$

由于 $\mathrm{Re}(z)<0$, 容易知道 $\mathrm{Re}(a_m)<0\,(m=1,2,\cdots,s)$, 于是根据归纳法, 对于 $m=s,s-1,\cdots,1$ 有 $\mathrm{Re}(A_m)<0$. 因为 $R(z)=1+2/(-1+A_1)$, 如果 $A_1=-x+iy$, 那么

$$\begin{aligned}|R(z)|^2 &= |1+2/(-1+A_1)|^2\\ &= \left|\frac{1-x+iy}{-1-x+iy}\right|^2\\ &= \frac{(1-x)^2+y^2}{(1+x)^2+y^2}\end{aligned}$$

因为 $x>0$, 显然, $|R(z)|^2<1$. 所以隐式 Runge-Kutta 方法是 A 稳定的. 顺便指出,上述引理 9.1 中当 $d=1,2$ 时, $R(z)=\frac{1}{2}N_s\Big/\frac{1}{2}D_s$ 为第一、二下对角 Padé 近似,因该式的分母的次数高于分子的次数, 所以, 若 $\mathrm{Re}(z)<0$ 且当 $\mathrm{Re}(z)\to-\infty$ 时, 有 $|R(z)|\to 0$. 因此在这种情形下这个方法是 L 稳定的.

§3 隐式 Runge-Kutta 方法的其他稳定性

上节讨论的方法的 A 稳定性, 是针对线性常系数的试验方程 $y'=\lambda y$ 来讨论的. 本节将考虑更加一般的非线性试验方程, 引

进相应的稳定性概念，并讨论方法的稳定性准则.

1975 年， Butcher[34] 首先对自守系统引进了 B 稳定的概念. 考虑试验方程组

$$y' = f(y), \ f: R^N \to R^N, \tag{9.28}$$

我们用到单调性条件或压缩性条件，即

$$\langle f(y) - f(z), y - z \rangle \leqslant 0, \ \text{对所有的} \ y, z \in R^N, \tag{9.29}$$

其中 $\langle \cdot , \cdot \rangle$ 表示 R^N 上的内积，R^N 表示N维实空间，$\|\cdot\|$ 表示对应的模。

设 (9.28) 中的 f 满足单调性条件，且 $\{y_n\}$ 和 $\{z_n\}$ 是由隐式 Runge-Kutta 公式 (9.2)、(9.3) 用同样的步长 h 求解同一个微分方程组 (9.28) 得到的两个近似解序列，

$$y_n = y_{n-1} + h \sum_{i=1}^{s} b_i f(Y_i), \tag{9.30}$$

$$z_n = z_{n-1} + h \sum_{j=1}^{s} b_j f(Z_j), \tag{9.31}$$

其中 $Y_1, Y_2, \cdots, Y_s$ 和 $Z_1, Z_2, \cdots, Z_s$ 分别表示从 y_{n-1} 计算 y_n 和从 z_{n-1} 计算 z_n 的中间结果，即

$$Y_i = y_{n-1} + h \sum_{j=1}^{s} a_{ij} f(Y_j), \tag{9.32}$$

$$Z_i = z_{n-1} + h \sum_{j=1}^{s} a_{ij} f(Z_j), \ i = 1, \cdots, s. \tag{9.33}$$

如果我们用 A 表示元素为 $a_{ij}(i, j = 1, \cdots, s)$ 的矩阵，利用 $a_{ij}^{(m)}$ 表示 A^m 的元素. 如果 A 是非奇异的，用 $a_{ij}^{(-1)}$ 来表示逆矩阵 A^{-1} 的元素. 这样，我们可以将 (9.32)、(9.33)、(9.30)、(9.31) 分别写成

$$hf(Y_i) = \sum_{j=1}^{s} a_{ij}^{(-1)}(Y_j - y_{n-1}), \tag{9.34}$$

$$hf(Z_i) = \sum_{j=1}^{s} a_{ij}^{(-1)}(Z_j - z_{n-1}), \tag{9.35}$$

$$y_n = y_{n-1} + \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)}(Y_j - y_{n-1}), \tag{9.36}$$

$$z_n = z_{n-1} + \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)}(Z_j - z_{n-1}). \tag{9.37}$$

定义 9.1 隐式 Runge-Kutta 方法称为B稳定的，如果对于所有满足条件(9.29)的自守系统(9.28)的数值解序列$\{y_n\}$，$\{z_n\}$，都有 $\|y_n - z_n\| \leqslant \|y_{n-1} - z_{n-1}\|$.

显然，由B稳定性可以推出A稳定性．设一个方法是B稳定的，将此方法用于试验方程 $y' = \lambda y(\mathrm{Re}(\lambda) < 0)$，则得 $y_n = R(h\lambda)y_{n-1}$. 试验方程的 $f = \lambda y$ 满足单调性条件，由

$$\|y_n - z_n\| \leqslant \|y_{n-1} - z_{n-1}\|$$

可以得出 $|R(h\lambda)| \leqslant 1$，即方法是$A$稳定的．（这里包括了 $|R(h\lambda)| = 1$）．例如，令方程组的f是一个常数矩阵

$$\begin{pmatrix} \lambda & -\mu \\ \mu & \lambda \end{pmatrix},$$

其特征值为 $\lambda \pm i\mu$. 如果 $\lambda \leqslant 0$，利用通常的内积则有 $\langle f(y) - f(z), y - z\rangle = \lambda\|y - z\|^2 \leqslant 0$，即满足单调性条件.

下面给出 Runge-Kutta 方法为B稳定的充分条件.

定理 9.3 如果隐式 Runge-Kutta 方法满足

1）A是非奇异的.

2）$b_1, b_2, \cdots, b_s$ 是非负的.

3）$c_1, c_2, \cdots, c_s$ 是互异的节点.

4）$0 \leqslant \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)} \leqslant 2$.

5）阶至少是 $2s - 2$.

6）$\sum_{j=1}^{s} a_{ij}c_j^{m-1} = c_i^m/m,\ m = 1, 2, \cdots, s-1, i = 1, \cdots, s$,

7）$\sum_{i=1}^{s} b_i c_i^{l-1} a_{ij} = b_j(1 - c_j^l)/l,\ l = 1, \cdots, s-1$,

$$j = 1, 2, \cdots, s.$$

则方法是 B 稳定的.

证明　若将 (9.30) 两端在 t_{n-1} 展开，比较 $h^k y_{n-1}^{(k)}$ 的系数，由 5) 得

$$\sum_{j=1}^{s} b_j c_j^{k-1} = \frac{1}{k}, \quad k = 1, \cdots, 2s-2. \qquad (9.37_1)$$

令 $\beta_{lm} = \sum_{i,j=1}^{s} b_i c_i^l a_{ij}^{(-1)} c_j^m \ (l, m = 0, 1, \cdots, s-1)$. 由 4) 得 $0 \leqslant \beta_{00} \leqslant 2$. 如果 $1 \leqslant m \leqslant s-1$，那么由 6) 和 (9.37_1)，我们有

$$\begin{aligned}\beta_{lm} &= m \sum_{i,j,k=1}^{s} b_i c_i^l a_{ij}^{(-1)} a_{jk} c_k^{m-1} \\ &= m \sum_{i=1}^{s} b_i c_i^{l+m-1} = m/(l+m).\end{aligned}$$

如果 $1 \leqslant l \leqslant s-1$，应用 7)、5) 和 (9.37_1)，有

$$\begin{aligned}\beta_{00} - \beta_{l0} &= \sum_{i,j=1}^{s} b_i (1 - c_i^l) a_{ij}^{(-1)} \\ &= l \sum_{i,j,k=1}^{s} b_k c_k^{l-1} a_{ki} a_{ij}^{(-1)} = l/l = 1.\end{aligned}$$

因此，$\beta_{l0} = \beta_{00} - 1$. 现在我们来计算 $\beta_{lm} + \beta_{ml} - \beta_{0l}\beta_{0m}$. 当 $l, m = 1, 2, \cdots, s-1$ 时，它的值是 $\dfrac{m}{l+m} + \dfrac{l}{m+l} - 1 = 0$. 当 $l, m = 0$ 时，它的值是 $\beta_{00}(2 - \beta_{00}) \geqslant 0$.

令 (i, j) 元素为

$$b_i a_{ij}^{(-1)} + b_j a_{ji}^{(-1)} - \sum_{k=1}^{s} b_k a_{ki}^{(-1)} \sum_{k=1}^{s} b_k a_{kj}^{(-1)} \quad (i, j = 1, 2, \cdots, s)$$

的矩阵为 M，而 C 是 (i, l) 元素为 $c_i^l (i = 1, \cdots, s, \ l = 0, 1, \cdots, s-1)$ 的矩阵. $C^T M C$ 的 (l, m) 元素是 $\beta_{lm} + \beta_{ml} - \beta_{0l}\beta_{0m}$，所以，$M$ 是秩为 1 或 0 的非负定的二次型矩阵. 如果 $\xi_0, \xi_1, \cdots, \xi_s \in R^N$，设 $M = (m_{ij})$，$\xi_i = (\xi_{i1}, \xi_{i2}, \cdots, \xi_{iN})^T$，则有

$$\sum_{i,j=1}^{s} m_{ij}\langle \xi_i, \xi_j\rangle = \left\langle \sum_{i,j=1}^{s} m_{ij}\xi_i, \xi_j \right\rangle$$

$$= \sum_{k=1}^{N}\left(\sum_{i,j=1}^{s} m_{ij}\xi_{ik}\xi_{jk}\right) \geqslant 0,$$

即是

$$\sum_{i,j=1}^{s}\left(b_i a_{ij}^{(-1)} + b_j a_{ji}^{(-1)} - \sum_{k=1}^{s} b_k a_{ki}^{(-1)} \sum_{k=1}^{s} b_k a_{kj}^{(-1)}\right)\langle \xi_i, \xi_j\rangle \geqslant 0.$$

由于

$$\sum_{i,j=1}^{s} (b_i a_{ij}^{(-1)} + b_j a_{ji}^{(-1)})\langle \xi_i, \xi_j\rangle$$

$$= \sum_{i=1}^{s} 2b_i \left\langle \sum_{j=1}^{s} a_{ij}^{(-1)}\xi_j, \xi_i\right\rangle$$

$$= \sum_{i=1}^{s} 2b_i \left\langle \sum_{j=1}^{s} a_{ij}^{(-1)}\xi_j, \xi_0 + \xi_i\right\rangle$$

$$- 2\left\langle \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)}\xi_j, \xi_0\right\rangle$$

和

$$\sum_{i,j=1}^{s}\left(- \sum_{k=1}^{s} b_k a_{ki}^{(-1)} \sum_{k=1}^{s} b_k a_{kj}^{(-1)}\right)\langle \xi_i, \xi_j\rangle$$

$$= -\left\|\sum_{i,j=1}^{k} b_i a_{ij}^{(-1)}\xi_j\right\|$$

$$= -\left\|\sum_{i,j=1}^{k} b_i a_{ij}^{(-1)}\xi_j + \xi_0\right\|^2 + \|\xi_0\|^2$$

$$- 2\left\langle \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)}\xi_j, \xi_0\right\rangle,$$

故上式可写成

$$\sum_{i=1}^{s} 2b_i \left\langle \sum_{j=1}^{s} a_{ij}^{(-1)}\xi_j, \xi_0 + \xi_i\right\rangle + \|\xi_0\|^2$$

$$- \left\| \xi_0 + \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)}\xi_j\right\|^2 \geqslant 0. \qquad (9.38)$$

选取向量 $\xi_0 = y_{n-1} - z_{n-1}$, $\xi_i = (Y_i - y_{n-1}) - (Z_i - z_{n-1})$, $(i = 1, \cdots, s)$，所以由 (9.34)、(9.35)，并用单调性条件，可得

$$\left\langle \sum_{j=1}^{s} a_{ij}^{(-1)} \xi_j, \xi_0 + \xi_i \right\rangle = h\langle f(Y_i) - f(Z_i), Y_i - Z_i \rangle \leqslant 0. \tag{9.39}$$

由于 $b_i \geqslant 0$, $(i = 1, 2, \cdots, s)$，结合 (9.38)、(9.39)，可以得到

$$\|\xi_0\|^2 - \|\xi_0 + \sum_{i,j=1}^{s} b_i a_{ij}^{(-1)} \xi_j\|^2 \geqslant 0,$$

再由 (9.36)、(9.37) 可以推出

$$\|y_n - z_n\| \leqslant \|y_{n-1} - z_{n-1}\|.$$

所以，方法是 B 稳定的. 定理证毕.

关于 B 稳定的充分条件，[34]、[39] 给出了一个更加简洁的定理. 我们这里只叙述定理，而不加证明.

定理 9.4 如果隐式 Runge-Kutta 方法使得有 $b_i \geqslant 0$ $(i = 1, \cdots, s)$，A 是非奇异的，且二次型 $\bar{Q}(\xi_1, \xi_2, \cdots, \xi_s) = \sum_{i,j=1}^{s} \bar{m}_{ij} \xi_i \xi_j$ 是非负定的，其中 $\bar{m}_{ij} = b_i a_{ij}^{(-1)} + b_j a_{ji}^{(-1)} - \sum_{k=1}^{s} b_k a_{ki}^{(-1)} \sum_{k=1}^{s} b_k a_{kj}^{(-1)}$ $(i, j = 1, 2, \cdots, s)$，$a_{ij}^{(-1)}$ 是 A^{-1} 的 (i, j) 元素. 那么 Runge-Kutta 方法是 B 稳定的.

1979 年，Burrage 和 Butcher[39] 把 B 稳定性概念推广到非自守系统. 引进了 BN 稳定性概念. 现在我们来讨论这个问题.

首先假设非自守系统

$$y' = f(t, y), \quad f: R^{N+1} \to R^N \tag{9.40}$$

的 f 对所有的 y、$z \in R^N$，$t \in R^1$ 满足

$$\langle f(t, y) - f(t, z), y - z \rangle \leqslant 0. \tag{9.41}$$

定义 9.2 隐式 Runge-Kutta 方法称为 BN 稳定的，如果对所有的满足条件 (9.41) 的非自守系统 (9.40) 的两个数值解序列 $\{y_n\}$ 和 $\{z_n\}$，那么 $\|y_n - z_n\| \leqslant \|y_{n-1} - z_{n-1}\|$ 成立.

Burrage 和 Butcher 给出了 B 稳定和 BN 稳定的新的准则．这个准则与二次型 $Q(\xi_1,\xi_2,\cdots,\xi_s)=\sum_{i,j=1}^{s} m_{ij}\xi_i\xi_j$ 相联系，其中

$$m_{ij}=b_i a_{ij}+b_j a_{ji}-b_i b_j. \tag{9.42}$$

由此，研究 B 稳定性只要验证这个代数性质是否成立．

定理 9.5 如果隐式 Runge-Kutta 方法的权系数 $b_1, b_2, \cdots, b_s \geqslant 0$，且二次型 Q 是非负定的，则该方法是 BN 稳定的．

证明 令 $v_0=y_{n-1}-z_{n-1}$，$v_j=Y_j-Z_j$，$w_j=hf(t_{n-1}+hc_j, Y_j)-hf(t_{n-1}+hc_j, Z_j)(j=1, 2, \cdots, s)$ 和 $v=y_n-z_n$．于是，由类似于 (9.32)、(9.33) 的公式得

$$v_i=v_0+\sum_{j=1}^{s} a_{ij}w_j,\ i=1, 2, \cdots, s, \tag{9.43}$$

并由类似于 (9.30)、(9.31) 的公式得

$$v=v_0+\sum_{j=1}^{s} b_j w_j. \tag{9.44}$$

v 的模的平方是

$$\|v\|^2=\|v_0\|^2+2\sum_{i=1}^{s} b_i\langle v_0, w_i\rangle + \sum_{i=1}^{s}\sum_{j=1}^{s} b_i b_j\langle w_i, w_j\rangle. \tag{9.45}$$

由 (9.43) 和 w_i 的内积，我们有

$$\langle v_0, w_i\rangle=\langle v_i, w_i\rangle-\sum_{j=1}^{s} a_{ij}\langle w_i, w_j\rangle,$$

将其代入 (9.45) 得到

$$\|v\|^2=\|v_0\|^2+2\sum_{j=1}^{s} b_j\langle v_j, w_j\rangle-\sum_{i,j=1}^{s} m_{ij}\langle w_i, w_j\rangle.$$

设 $w_i=(w_{i1},\cdots, w_{iN})^T$，

$$\sum_{i,j=1}^{s} m_{ij}\langle w_i, w_j\rangle=\sum_{k=1}^{N}\left(\sum_{i,j=1}^{s} m_{ij}w_{ik}w_{jk}\right)\geqslant 0,$$

如果

$$\langle v_j, w_j\rangle = h\langle Y_j - Z_j, f(t_{n-1} + hc_j, Y_j) - f(t_{n-1} + hc_j, Z_j)\rangle \leqslant 0,$$

那么应用定理的假定条件，我们有

$$\|v\|^2 \leqslant \|v_0\|^2,$$

故该方法是 BN 稳定的. 定理证毕.

显然，定理 9.5 所叙述的条件也是 B 稳定的充分条件，即有如下推论:

推论 如果隐式 Runge-Kutta 方法满足定理 9.5 的条件，则该方法是 B 稳定的.

现在我们介绍 Burrage 和 Butcher[39] 提出的代数稳定性概念. 用 M 表示其 (i, j) 元素为 $b_i a_{ij} + b_j a_{ji} - b_i b_j$ 的对称矩阵.

定义 9.3 隐式 Runge-Kutta 方法称为代数稳定的，如果 $b_i \geqslant 0\ (i = 1, 2, \cdots, s)$，$\xi^T M \xi$ 是非负定的.

由定义我们知道代数稳定的方法一定是 BN 稳定的. A 稳定性只是讨论解线性常系数微分方程组时的数值稳定性，把这个概念推广到更广的情形就是所谓的 AN 稳定性.

下面我们来介绍这个概念，并且讨论它与 BN 稳定性的关系.

考虑试验方程

$$y' = q(t)y, \tag{9.46}$$

其中 q 是定义在实轴上的复值连续函数（为了保证问题是适定的），又假定 q 只取非正的实部值(为了保证它有一个在量值上不增加的解). (9.46) 虽然只是线性变系数的形式，但也反应了解非线性问题误差的变化性质.

用 Runge-Kutta 方法以步长 $h = t_n - t_{n-1}$，从 t_{n-1} 到 t_n 数值求解的计算过程中，将产生 q 的一些值，用 $\zeta_1, \zeta_2, \cdots, \zeta_s$ 来表示这些值与 h 的乘积. 因此，定义

$$\zeta_i = hq(t_{n-1} + hc_i)\ \ (i = 1, 2, \cdots, s), \tag{9.47}$$

并记 $\zeta = \mathrm{diag}(\zeta_1, \zeta_2, \cdots, \zeta_s)$. 若将 (9.2)、(9.3) 中的 hK_i 写成

K_i，其中的 n 取成 $n-1$．熟知，如果对于一个常系数线性系统，令 $z=hq$，则数值解 y_n 满足

$$y_n = R(z)y_{n-1}, \tag{9.47$_1$}$$

其中

$$R(z) = 1 + zb^T(I - zA)^{-1}e,$$

这儿 $e=(1,1,\cdots,1)^T$．容易证明，利用我们对非自守系统的推广，(9.47$_1$) 换成为

$$y_n = K(\zeta)y_{n-1}, \tag{9.48}$$

其中

$$K(\zeta) = 1 + b^T\zeta(I - A\zeta)^{-1}e. \tag{9.49}$$

一般来说，$K(\zeta)$ 是 $\zeta_1,\zeta_2,\cdots,\zeta_s$ 的有理函数，像上面一样，我们称 $K(\zeta)$ 为传播函数．现在我们定义 AN 稳定性，即对非自守系统的 A 稳定性．我们取 ζ_i 为任意的复数，$\mathrm{Re}(\zeta_i)\leqslant 0\,(i=1,2,\cdots,s)$，只要 Runge-Kutta 方法的 $c_i=c_j$ 就有 $\zeta_i=\zeta_j$ $(i,j=1,2,\cdots,s)$．

定义 9.4 隐式 Runge-Kutta 方法称为 AN 稳定的，如果对所有上述的 $\zeta=\mathrm{diag}(\zeta_1,\cdots,\zeta_s)$，方法的传播函数都满足

$$|K(\zeta)|\leqslant 1.$$

我们引进如下的定理，而不给予证明．

定理 9.6 如果隐式 Runge-Kutta 方法是 AN 稳定的，那么该方法也是 A 稳定的．

但是，该定理的逆命题是不成立的．例如考虑 Runge-Kutta 方法

$$\begin{array}{c|cc} \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\[4pt] \frac{3}{4} & \frac{3}{8} & \frac{3}{8} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array},$$

我们可以得到 $R(z)=(2+z)/(2-z)$，对于 $\mathrm{Re}(z)<0$，有 $|R(z)|<1$，于是该方法是 A 稳定的．对于传播函数 $K(\zeta)=$

$(8+3\zeta_1+\zeta_2)/(8-\zeta_1-3\zeta_2)$，显然，当 $\mathrm{Re}(\zeta_1)\leqslant 0$，$\mathrm{Re}(\zeta_2)\leqslant 0$ 时，$|K(\zeta)|\leqslant 1$，甚至不是有界的．所以方法不是 AN 稳定的．

因此，AN 稳定性比 A 稳定性的要求要多一些．对于 A 稳定性没有简单的代数准则来刻划．然而对 AN 稳定性却可以给出简单的代数描述．下面的定理 9.7 给出这种代数条件．为此，我们首先证明下面的引理．

引理 9.4 设 ζ 满足 $\det(I-A\zeta)\neq 0$，又令 $u=(I-A\zeta)^{-1}e$，则有

$$|K(\zeta)|^2-1=2\sum_{i=1}^{s}b_i\mathrm{Re}(\zeta_i)|u_i|^2-\sum_{i,j=1}^{s}m_{i,j}\zeta_i\bar{u}_i\zeta_j u_j, \quad (9.50)$$

其中 m_{ij} 由 (9.42) 给出．

证明　由于 $u_i=1+\sum_{j=1}^{s}a_{ij}\zeta_j u_j(i=1,2,\cdots,s)$，用 b_i 乘之，交换 i 和 j，并取共轭，我们有

$$b_i=b_iu_i-\sum_{j=1}^{s}b_ia_{ij}\zeta_ju_j\ (i=1,2,\cdots,s), \quad (9.51)$$

$$b_j=b_j\bar{u}_j-\sum_{i=1}^{s}b_ja_{ji}\zeta_i\bar{u}_i\ (i=1,2,\cdots,s). \quad (9.52)$$

现在我们用 $K(\zeta)=1+\sum_{j=1}^{s}b_j\zeta_ju_j$ 的共轭（且以 i 代替 j）来乘 $K(\zeta)$，得到

$$|K(\zeta)|^2-1=\sum_{i=1}^{s}b_i\zeta_i\bar{u}_i+\sum_{j=1}^{s}b_j\zeta_ju_j+\sum_{i,j=1}^{s}b_ib_j\zeta_iu_i\zeta_j\bar{u}_j,$$

并且将 (9.51)、(9.52) 均代入上式右边的前两项，引理就得证．

定理 9.7 代数稳定的 Runge-Kutta 方法是 AN 稳定的，反之，若 $c_1,c_2,\cdots,c_s$ 是互异的，则 AN 稳定的方法也是代数稳定的．

证明　如果方法是代数稳定的，由定理 9.5 可知方法是 BN 稳定的．于是立即可得定理的第一部分．另外，从 (9.50) 也可以

得到所要的结果，由于在一个代数稳定的方法中，右边部分一定是非正的.

现在来证明定理的第二部分. 首先我们注意，由于 $c_1, c_2, \cdots, c_s$ 是互异的，$\zeta_1, \zeta_2, \cdots, \zeta_s$ 可以在复平面上任意选取. 为了证明 $b_i \geqslant 0$，我们用反证法，假定有某个 $b_i < 0$，并选取 $\zeta_i = -\varepsilon$，其中 ε 为一个小的正实数，并且对于 $j \neq i$，$\zeta_j = 0$. 由于 $u = (I - A\zeta)^{-1}e$，可解出 $u_k = \dfrac{1 + \varepsilon(a_{ii} - a_{ik})}{1 + \varepsilon a_{ii}}$，特别

$$u_i = \frac{1}{1 + \varepsilon a_{ii}},$$

故可选取 ε 这样小，使得 $\frac{1}{2} < u_i < \frac{3}{2}$，(9.50) 的右边部分变成

$$-2b_i\varepsilon u_i^2 - (2b_i a_{ii} - b_i^2)\varepsilon^2 u_i^2,$$

故若 $b_i < 0$，对于充分小的 ε，上式是正的，与 AN 稳定的方法的 $|K(\zeta)| \leqslant 1$ 矛盾. 因此 b_i 不能为负.

为了证明系数为 m_{ij} 的二次型的非负定性质，我们仍用反证法. 对于一个 AN 稳定的方法，设有实数组 $\xi_1, \xi_2, \cdots, \xi_s$ 使得 $\sum_{i,j=1}^{s} m_{ij}\xi_i\xi_j < 0$，则选取 $\zeta_1, \zeta_2, \cdots, \zeta_s$ 为纯虚数组 $\varepsilon\xi_1 i, \varepsilon\xi_2 i, \cdots, \varepsilon\xi_s i$；其中 ε 为小的正数. 令 $u_i = 1 + \varphi_i(\varepsilon)$，其中当 $\varepsilon \to 0$ 时，$|\varphi_i(\varepsilon)| \to 0$. 计算 (9.50) 的右边部分，得到

$$|K(\zeta)|^2 - 1 = -\varepsilon^2 \sum_{i,j=1}^{s} m_{ij}\xi_i\xi_j + \varepsilon^2 c(\varepsilon)$$

其中当 $\varepsilon \to 0$ 时，$|c(\varepsilon)| \to 0$. 于是，对于充分小的 ε，$|K(\zeta)| > 1$. 这一事实与 AN 稳定的假定矛盾. 故二次型 $\sum_{i,j=1}^{s} m_{ij}\xi_i\xi_j$ 不能为负定的. 总之，方法是代数稳定的. 定理证毕.

将这个结果与定理 9.5 结合起来，我们立即可以得到如下定理.

定理 9.8 具有 $c_1, c_2, \cdots, c_s$ 互异的隐式 Runge-Kutta 方法为 AN 稳定的充要条件是该方法为 BN 稳定的.

这个结果有重要的意义，因为它将非线性问题的稳定性质与线性问题的稳定性质建立了联系，并用一定的代数准则来刻划. 这些条件之间的联结与所考虑的试验方程的非自守性质有关. 但是，在 A 稳定性与 B 稳定性之间尚未发现存在这样的充分和必要条件.

在 [34] 中，已证明一些高阶的方法类是 B 稳定的，于是自然要考虑它们是不是代数稳定的. 令 $\overline{M}=(\overline{m}_{ij})$ 是一个矩阵，$\overline{m}_{ij}$ 为定理 9.4 中的 $\overline{m}_{ij}$. 容易知道，$A^T\overline{M}A=M$. 如果 A 是非奇异的，我们可以叙述下面的定理.

定理 9.9 如果 A 是非奇异的，$b_i\geqslant 0(i=1,2,\cdots,s)$，又 $\overline{M}$ 是非负定的，那么 Runge-Kutta 方法是代数稳定的.

下面我们介绍代数稳定的方法的几个例子.

例 9.1 二级二阶隐式 Runge-Kutta 方法为

$$\begin{array}{c|cc} \lambda & \lambda & 0 \\ 1-\lambda & 1-2\lambda & \lambda \\ \hline & 1/2 & 1/2 \end{array},$$

如果 $\lambda\geqslant\frac{1}{4}$，方法是代数稳定的. 由于

$$M=\left(\lambda-\frac{1}{4}\right)\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

又，如果 $\lambda=(3+\sqrt{3})/6$，方法是代数稳定的，且其阶为三.

例 9.2 Nφrsett[90] 构造了一类三级四阶方法

$$\begin{array}{c|ccc} \lambda & \lambda & & \\ \frac{1}{2} & \frac{1}{2}-\lambda & \lambda & \\ 1-\lambda & 2\lambda & 1-4\lambda & \lambda \\ \hline & b_1 & b_2 & b_3 \end{array},$$

其中

$$b_1=b_3=1/6(1-2\lambda)^2,\quad b_2=1-2b_1,$$
$$1/24-\lambda/2+3\lambda^2/2-\lambda^3=0.$$

我们注意到上面的最后一个方程的零点近似等于

$$\lambda_1 = 1.06858,\ \lambda_2 = 0.30254,\ \lambda_3 = 0.12889.$$

现在构造对应于这个方法的矩阵 M，但是首先注意，对于 $\lambda \in ((3-\sqrt{3})/6,\ (3+\sqrt{3})/6)$，$b_2 < 0$。因此，如果我们希望得到代数稳定性，$\lambda$ 不能等于 λ_2。因为 M 是对称的，又 $b_1 = b_3$，$m_{11} = m_{33} = m_{13} = m_{31}$。

另外，因为 $24\lambda^3 - 36\lambda^2 + 12\lambda - 1 = 0$，我们有 $(1-4\lambda)/6(1-2\lambda)^2 - (1-2\lambda)(3(1-2\lambda)^2 - 1)/6(1-2\lambda)^2 = 0$，当 $b_1 = b_3$ 时，有 $m_{23} = m_{21}$，还可证明 $4m_{11} = -2m_{12} = m_{22}$，其中 $m_{11} = b_1(2\lambda - b_1)$，$m_{12} = b_2(1/2 - \lambda - b_2)$，$m_{22} = b_2(2\lambda - b_2)$，于是

$$M = b_1(2\lambda - 1/6(1-2\lambda)^2)\begin{pmatrix} 1 & -2 & 1 \\ -2 & 4 & -2 \\ 1 & -2 & 1 \end{pmatrix}.$$

它只当 $\lambda = \lambda_1$ 时是非负定的，所以如果 $\lambda \simeq 1.06858$，本例给出的方法是代数稳定的。

正如例子中所看到的，对于更高阶 $(s \geqslant 4)$ 的隐式 Runge-Kutta 方法代数稳定性的检验可能变得更加繁杂，这里就不讨论了。

本 章 附 注

§1 主要材料取自 Butcher 的 [33]，Ehle 的 [50]，Axelsson 的 [23]，Lapidus 等人的书 [71]，汤怀民的 [12]。

§2 的材料主要取自 Butcher 的 [36]。

§3 的材料主要取自 Butcher 的 [34]，Burrage 和 Butcher 的 [39] 以及汤怀民的 [12]。

第十章　隐式 Runge-Kutta 方法的实现

从前一章我们看到,隐式 Runge-Kutta 公式是高阶的, A 稳定的数值积分方法. 但是,应用隐式 Runge-Kutta 公式,除了求解线性微分方程组，一般存在着较大的困难. 当我们使用 s 级的隐式 Runge-Kutta 方法求解 m 个方程的一阶常微分方程组时,每积分一个步长，需要求解 sm 个联立的非线性代数方程组. 如果这些代数方程组是用 Newton 方法或者它的变形来求解时，则在每迭代一步必须求解 sm 个线性代数方程组. 除非这个方程组是稀疏的,并且应用了稀疏矩阵的技术外,每迭代一步需要的乘法个数的数量级为 s^3m^3. 于是,隐式 Runge-Kutta 方法的实用性就受到限制. 本章将介绍为克服这一困难，对隐式 Runge-Kutta 方法的实现所采取的一些办法. 主要内容包括：等效代换的迭代方法；修改的 Newton 迭代方法；对角线隐式 Runge-Kutta 方法；Butcher 矩阵变换及其相应的方法；Rosenbrock 半隐式 Runge-Kutta 方法以及广义 Runge-Kutta 方法.

应当指出,随着计算技术的不断发展,特别是随着每秒亿万次计算速度的向量计算机的出现,高阶的 A 稳定的隐式 Runge-Kutta 方法就可能用并行算法来实现.

§1　等效代换的迭代方法

为了实现隐式 Runge-Kutta 方法，1973 年，Chipman[48] 提出了一个等效代换的迭代方法. 主要思想是，做一个等价的新的微分方程组代替原来的微分方程组. 在这个基础上得到一个迭代格式,并且可以得到这个格式收敛的充分条件.

我们考虑数值求解初值问题

$$y' = f(t, y), \quad y(0) = y_0, \tag{10.1}$$

其中 f 和 y 均是 m 维向量．应用 s 级的隐式 Runge-Kutta 方法

$$y_{n+1} = y_n + hWK, \quad n = 0, 1, 2, \cdots, \tag{10.2}$$

这里

$$K = \begin{pmatrix} K_1 \\ \vdots \\ K_s \end{pmatrix}, \quad \begin{array}{l} K_i = f\left(t_n + c_i h, y_n + h \sum_{j=1}^{s} b_{ij} K_j\right), \\ i = 1, 2, \cdots, s, \end{array} \tag{10.3}$$

其中 K_i 是 m 个元素的列块，K 是 $m \times s$ 维列向量．又，

$$W = (W_1 I, \cdots, W_s I),$$

I 是 m 阶单位矩阵．

现假定 f 对 y 的一阶偏导数存在，将 (10.1) 改写成

$$y' = Ay + r(t, y), \quad y(0) = y_0, \tag{10.4}$$

其中 A 是 m 阶常矩阵，开始时可选成为 $(\partial f/\partial y)_{(t_0, y_0)}$ 以及

$$r(t, y) = f(t, y) - Ay.$$

下面我们正式定义前一章已使用过的不同阶的两个矩阵之间的一种运算．

定义 10.1 若 s 阶矩阵 $B = (b_{ij})$，m 阶矩阵 $A = (a_{ij})$，有

$$B \otimes A = \begin{pmatrix} b_{11}A & \cdots & b_{1s}A \\ \vdots & & \vdots \\ b_{s1}A & \cdots & b_{ss}A \end{pmatrix},$$

则称这种运算是 B 与 A 的 Kronecker 积．

那么就得到关于 K 的隐式方程

$$K = \begin{pmatrix} A \\ \vdots \\ A \end{pmatrix} y_n + h(B \otimes A)K + \begin{pmatrix} r\left(t_n + c_1 h, y_n + h \sum_{j=1}^{s} b_{1j} K_j\right) \\ \vdots \\ r\left(t_n + c_s h, y_n + h \sum_{j=1}^{s} b_{sj} K_j\right) \end{pmatrix}. \tag{10.5}$$

对于$m\times s$个未知数K的非线性方程组(10.5)迭代求解。因此，我们考虑由

$$K^{(q+1)}=(E-h(B\otimes A))^{-1}\left(\begin{pmatrix}A\\ \vdots\\ A\end{pmatrix}y_n+\begin{pmatrix}r\left(t_n+c_1h,\,y_n+h\sum_{j=1}^{s}b_{1j}K_j^{(q)}\right)\\ \vdots\\ r\left(t_n+c_sh,\,y_n+h\sum_{j=1}^{s}b_{sj}K_j^{(q)}\right)\end{pmatrix}\right)$$

来定义迭代。这里E为$m\times s$阶单位矩阵。应用$r(t,y)=f(t,y)-Ay$，并作变换(I是m阶单位矩阵)

$$L=\begin{pmatrix}L_1\\ \vdots\\ L_s\end{pmatrix}=\begin{pmatrix}b_{11}I & \cdots & b_{1s}I\\ \vdots & & \vdots\\ b_{s1}I & \cdots & b_{ss}I\end{pmatrix}\begin{pmatrix}K_1\\ \vdots\\ K_s\end{pmatrix}=(B\otimes I)K.$$

这样,最后的方程就变成

$$L^{(q+1)}=((P\otimes I)-\mathrm{diag}(hA))^{-1}\cdot\begin{pmatrix}f(t_n+c_1h,\,y_n+hL_1^{(q)})-hAL_1^{(q)}\\ \vdots\\ f(t_n+c_sh,\,y_n+hL_s^{(q)})-hAL_s^{(q)}\end{pmatrix},\qquad(10.6)$$

其中$P=B^{-1}$，并假设其存在.

格式(10.6)的收敛性的充分条件由下面的定理给出:

定理 10.1 如果$f(t,y)$对y具有连续的一阶偏导数，且对于$t\in(t_n,t_{n+1})$和$y=y_n+\theta(y_{n+1}-y_n)$，$0<\theta<1$，$\|(\partial f/\partial y)_{(t,y)}-A\|$充分小,那么由(10.6)定义的迭代格式收敛.

证明　我们应用向量模$\|y\|=\max\limits_i|y_i|$和相应的矩阵模

$$\|A\| = \max_i \sum_j |a_{ij}|.$$

令

$$\varphi(L) = ((P\otimes I) - \mathrm{diag}(hA))^{-1} \cdot \begin{pmatrix} f(t_n + c_1 h, y_n + hL_1) - hAL_1 \\ \vdots \\ f(t_n + c_s h, y_n + hL_s) - hAL_s \end{pmatrix},$$

于是对于(10.6)的逐次迭代L和$\bar{L}$，有

$$\begin{aligned}
&\|\varphi(L) - \varphi(\bar{L})\| \\
&\leqslant H\left\| \begin{pmatrix} f(t_n + c_1 h, y_n + hL_1) - f(t_n + c_1 h, y_n + h\bar{L}_1) \\ \vdots \\ f(t_n + c_s h, y_n + hL_s) - f(t_n + c_s h, y_n + h\bar{L}_s) \end{pmatrix} - h\begin{pmatrix} A(L_1 - \bar{L}_1) \\ \vdots \\ A(L_s - \bar{L}_s) \end{pmatrix} \right\| \\
&\leqslant H\left\| \begin{pmatrix} \dfrac{\partial f}{\partial y}(t_n + c_1 h, y_n + h\zeta_1)(L_1 - \bar{L}_1) \\ \vdots \\ \dfrac{\partial f}{\partial y}(t_n + c_s h, y_n + h\zeta_s)(L_s - \bar{L}_s) \end{pmatrix} - h\begin{pmatrix} A(L_1 - \bar{L}_1) \\ \vdots \\ A(L_s - \bar{L}_s) \end{pmatrix} \right\|,
\end{aligned}$$

其中

$$H = \|((P\otimes I) - \mathrm{diag}(hA))^{-1}\|,\quad \zeta_i = \bar{L}_i + \theta(L_i - \bar{L}_i),$$
$$0 < \theta < 1,\quad i = 1, \cdots, s.$$

上述不等式也可写成

$$\|\varphi(L)-\varphi(\bar{L})\| \leqslant H\left\|\frac{\partial f}{\partial y}(t_n+c_v h, y_n+h\zeta_v)-hA\right\| \cdot \|L-\bar{L}\|$$

对于 $1 \leqslant v \leqslant s$.

因此，当 $t \in (t_n, t_{n+1})$ 和 $y = y_n + \theta(y_{n+1} - y_n)$，$0 < \theta < 1$ 时，$\|(\partial f/\partial y)(t, y) - hA\| < 1/H$，则 φ 是一个压缩映象．定理证毕．

§2 修改的 Newton 迭代方法

应用 s 级隐式 Runge-Kutta 方法求解一阶 m 维常微分方程组，每积分一步要解 sm 个非线性代数方程组．1977 年，Bickart[28] 提出了一个求解该类型的非线性代数方程组的修改的 Newton 迭代方法，使得计算量大大减少．现在我们介绍这个方法．

像在本章第一节那样，我们考虑初值问题

$$y' = f(t, y), \quad y(0) = y_0, \tag{10.7}$$

其中 y 和 f 是 m 维向量．令 y_n 表示 $y(t_n)$ 的近似值，$t_n = t_0 + nh$．y_n 由 Runge-Kutta 方法来计算

$$y_n = y_{n-1} + h\sum_{i=1}^{s} b_i K_i, \tag{10.8}$$

其中

$$K_i = f\left(t_{n-1} + c_i h, y_{n-1} + h\sum_{j=1}^{s} a_{ij} K_j\right), \tag{10.9}$$
$$i = 1, \cdots, s.$$

令

$$P_i = y_{n-1} + h\sum_{j=1}^{s} a_{ij} K_j, \tag{10.10}$$
$$i = 1, \cdots, s.$$

或者等价地令

$$P_i = y_{n-1} + h\sum_{j=1}^{s} a_{ij} f_j \quad (i = 1, \cdots, s), \tag{10.11}$$

这里

$$f_i = f(t_{n-1} + c_i h, P_i). \tag{10.12}$$

注意，我们可以把 P_i 看成是 $y(t)$ 在 $t = t_{n-1} + c_i h \in [t_{n-1}, t_n]$ 处的近似. 假定矩阵 $A = (a_{ij})$ 是非奇异的，由 (10.11) 中解出 f_i，得

$$hf_i = \sum_{j=1}^{s} \alpha_{ij} P_j - \left(\sum_{j=1}^{s} \alpha_{ij}\right) y_{n-1}, \tag{10.13}$$
$$i = 1, \cdots, s,$$

其中 α_{ij} 是 A^{-1} 的第 i 行第 j 列的元素. 方程组 (10.13) 可以表示成矩阵的形式

$$(A^{-1} \otimes I_m)\hat{P} - h\hat{D} - (\mathrm{a} \otimes I_m) y_{n-1} = 0, \tag{10.14}$$

这里 I_m 是 m 阶单位矩阵，$A^{-1} = (\alpha_{ij})$，

$$\mathrm{a} = \begin{pmatrix} \sum\limits_{j=1}^{s} \alpha_{1j} \\ \vdots \\ \sum\limits_{j=1}^{s} \alpha_{sj} \end{pmatrix}, \quad \hat{P} = \begin{pmatrix} P_1 \\ \vdots \\ P_s \end{pmatrix}, \quad \hat{D} = \begin{pmatrix} f_1 \\ \vdots \\ f_s \end{pmatrix}.$$

由 (10.14) 对 $\hat{P}$ 建立 Newton 迭代格式. 设 $\hat{P}$ 的第 $(l+1)$ 次迭代记为 $\hat{P}^{l+1}$，得

$$((A^{-1} \otimes I_m) - hJ^l)(\hat{P}^{l+1} - \hat{P}^l)$$
$$= -((A^{-1} \otimes I_m)\hat{P}^l - h\hat{D}^l - (\mathrm{a} \otimes I_m) y_{n-1}), \tag{10.15}$$

由 (10.15) 可以解出 $\hat{P}^{l+1}$，其中

$$J^l = \mathrm{diag}((\partial f/\partial y)_{(t_{n-1}+c_1 h, P_1^l)}, \cdots,$$
$$(\partial f/\partial y)_{(t_{n-1}+c_s h, P_s^l)}) \tag{10.16}$$

是块对角矩阵. 我们知道，对 (10.11) 的等价方程组 (10.14)，应用 Newton 迭代法求 P_i ($\hat{P}$ 的元素)等价于对 (10.9) 应用 Newton 迭代法求 K_i (K 的元素).

如果 (10.16) 中的 J^l 用块对角矩阵 J 来代替，即用 f 的 Jacobi 矩阵来代替，

$$J = \mathrm{diag}(F_{\hat{n}}, \cdots, F_{\hat{n}}) = I_s \otimes F_{\hat{n}}, \qquad (10.17)$$

其中 I_s 是 s 阶单位矩阵，而

$$F_{\hat{n}} = f_y(t_{\hat{n}}, y_{\hat{n}}), \qquad (10.18)$$

不是在每次迭代中都求 J 的值，而只是在点 $t_{\hat{n}}$ 上求值，也即是在 $\hat{n} = 0, 1, \cdots, n-1$ 的积分步的节点上求值.

应用 (10.17) 到 (10.15)，我们得到修改的 Newton 法.

$$((A^{-1} \otimes I_m) - h(I_s F_{\hat{n}}))(\hat{P}^{l+1} - \hat{P}^l)$$
$$= -((A^{-1} \otimes I_m)\hat{P}^l - h\hat{D}^l - (\mathrm{a} \otimes I_m)y_{n-1}), \qquad (10.19)$$

这个方程等价于下面的矩阵方程

$$(hF_{\hat{n}})(P^{l+1} - P^l) - (P^{l+1} - P^l)(A^{-1})^T = E^l, \qquad (10.20)$$

其中

$$P^l = (P_1^l, \cdots, P_s^l), \quad E^l = P^l(A^{-1})^T - hD^l - y_{n-1}\mathrm{a}^T,$$
$$D^l = (f_1^l, \cdots, f_s^l),$$

$(A^{-1})^T$ 是 A^{-1} 的转置.

假定 $a(\lambda)$ 表示 $(A^{-1})^T$ 的特征多项式，即令

$$a(\lambda) = \det(\lambda I_s - (A^{-1})^T) = \sum_{k=0}^{s} a_k \lambda^{s-k}. \qquad (10.21)$$

如 Jameson (见 Bickart[28]) 所证明的，(10.20) 的解可以表示为

$$P^{l+1} - P^l = (a(hF_{\hat{n}}))^{-1}\left(\sum_{k=0}^{s-1} a_k M_{s-k}\right), \qquad (10.22)$$

其中

$$M_0 = 0, \quad M_1 = E^l,$$
$$M_k = (hF_{\hat{n}})M_{k-1} + M_{k-1}(A^{-1})^T - (hF_{\hat{n}})M_{k-2}(A^{-1})^T.$$

从上述过程中我们注意到，如 $\hat{n} < n-1$，那么 $(a(hF_{\hat{n}}))^{-1}$ 是在前面的时间步上计算的. 因此不需在当前的时间步上计算. 另一方面，如果 $\hat{n} = n-1$，那么 $(a(hF_{\hat{n}}))^{-1}$ 就需要求值. 因此需要一次 Jacobi 矩阵的计算和 $sm^3 + m^2$ 个数乘法. 数乘法的次数(包括除法的次数)、向量求值的次数和 Jacobi 矩阵计算的次数都列在下表. 第 n 个时间步的迭代步的步数用 l_n 来表示. 为了便于比较. 我们还在表 10.1 中列出了用 (10.19) 解 $\hat{P}^l$ 时应用

$((A^{-1}\otimes I_m) - h(I_s\otimes F_{\hat{n}}))$ 直接求逆所需要的乘法次数．称用 (10.19) 解 $\hat{P}$ 为旧方法，而用 (10.22) 解 P 为新方法．在新旧两种方法中，向量求值次数和 Jacobi 矩阵求值次数相同．

表 10.1　数乘法次数

$n-1$	新方法(对 P 求解)	旧方法(对 $\hat{P}$ 求解)	向量函数求值次数	Jacobi 矩阵求值次数
$>\hat{n}$	$sm+s$ $+l_n\{s^2m^2+(s^3+s^2)m\}$	$sm+s$ $+l_n\{s^2m^2+(s^2+s)m\}$	l_ns	0
$=\hat{n}$	sm^3+m^2+sm+s $+l_n\{s^2m^2+(s^3+s^2)m\}$	$s^3m^3+m^2+sm+s$ $+l_n\{s^2m^2+(s^2+s)m\}$	l_ns	1

从上表可以看出，对于大的 m，即求解的常微分方程组的方程个数很多和 $\hat{n}=n-1$ 时计算工作量最大．在一个时间步中对 P 求解的乘法的次数的表达式中主要项是 $sm^3+l_ns^2m^2$．而当直接求解 (10.19) 时，乘法次数的主要项是 $s^3m^3+l_ns^2m^2$．如果用 LU 分解而不是直接求逆，则主要项分别变成为 $(s-2/3)m^3+l_ns^2m^2$ 和 $(1/3)s^3m^3+l_ns^2m^2$．在实际求解的过程中看到，l_n 对新、旧两种方法实质上是一样的．顺便指出，新方法比旧方法所需要的数据存贮量小．

由 (10.22) 得知，新方法求逆是对 m 阶矩阵进行的，而不是对 sm 阶矩阵进行的．在许多常微分方程组的求解中，特别当 m 很大时，往往具有稀疏的 Jacobi 矩阵 $F_{\hat{n}}$，并且 $a(\lambda)$ 的次数 s 不太大．这时 $a(hF_{\hat{n}})$ 仍可能是非常稀疏的．应用稀疏矩阵技术仍有好处．例如 $F_{\hat{n}}$ 是带状矩阵，并有小的带宽就是这种情形．

§3　对角线隐式 Runge-Kutta 方法

我们知道用 Runge-Kutta 方法求解初值问题

$$y'=f(t,y),\ y(0)=y_0 \tag{10.23}$$

的近似值，其思想是通过求积分

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))\, dt \tag{10.24}$$

的近似值来得到从 t_n 到 t_{n+1} 的近似解，其中 $h = t_{n+1} - t_n$ 是当前的积分步长。我们选取求积节点 $c_1, c_2, \cdots, c_s$ 和权系数 $b_1, b_2, \cdots, b_s$，应用求积公式

$$y(t_{n+1}) = y(t_n) + h \sum_{i=1}^{s} b_i f(t_n + c_i h, y(t_n + c_i h)) + \text{误差项}, \tag{10.25}$$

记 $t_n + c_i h$ 为 $t_{n,i}$，$y(t_n)$ 的近似值为 y_n。在 (10.25) 中将 $y_{n,i}$ 的值代替 $y(t_{n,i})$ 的值。在同样的节点上数值求积得到 $y_{n,i}$

$$y_{n,i} = y_n + h \sum_{j=1}^{s} a_{ij} f(t_{n,j}, y_{n,j}), \quad i = 1, \cdots, s. \tag{10.26}$$

一般来说，这是一组隐式方程。求解 $y_{n,i}$ 并代入 (10.25) 中，得到 y 的下一个近似值

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i f(t_{n,i}, y_{n,i}), \tag{10.27}$$

与前面用 K 表达的 Runge-Kutta 公式一样，(10.26)，(10.27) 一起定义了 Runge-Kutta 方法。将其系数排成表

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \\ c_s & a_{s1} & \cdots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} \tag{10.28}$$

本章的前两节所介绍的方法，都是为了克服用 s 级隐式 Runge-Kutta 公式求解 m 个一阶微分方程组时，一般需要在每个时间步求解 ms 个方程的联立的非线性代数方程组的困难而设计的。在这一节中，我们从另外一个角度来研究隐式 Runge-Kutta 方法的实现，即基础于 Runge-Kutta 方法本身的改变，也就是在(10.28)中应用一个下三角形矩阵 (a_{ij})。Butcher 称其为半隐式 Runge-Kutta 方法。这就将解 ms 个方程的方程组分成解 s 次 m 个方程的方程组，即对 $i = 1, 2, \cdots, s$ 逐次求解(10.26)。若用 Newton

法迭代求解，则线性代数方程组的系数矩阵的形式为

$$I - ha_{ii}\partial f/\partial y,$$

如果在半隐式 Runge-Kutta 公式中所有的 a_{ii} 都相等，我们称它为对角线隐式 Runge-Kutta 公式，简称为 DIRK 公式.

现在介绍 Alexander 在[22]中所综述的一些结果. 对于 Runge-Kutta 公式，我们仍像前面那样作如下的约定：s 表示方法的级数，p 表示方法的阶数.

容易看出，隐式中点法则为 $(s, p) = (1, 2)$ 的 DIRK(1,2) 公式

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \tag{10.29}$$

是 A 稳定的.

在 Alexander 的综述中，指出 Crouzeix 已找出了所有二级三阶和三级四阶的半隐式 Runge-Kutta 方法，我们摘录如下：

定理 10.2 对于 $(s, p) = (2, 3)$ 和 $(s, p) = (3, 4)$，恰好存在 A 稳定的 DIRK 公式，即

$(s, p) = (2, 3)$

$$\begin{array}{c|cc} \frac{1}{2} + \frac{1}{2\sqrt{3}} & \frac{1}{2} + \frac{1}{2\sqrt{3}} & 0 \\ \frac{1}{2} - \frac{1}{2\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{2} + \frac{1}{2\sqrt{3}} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}, \tag{10.30}$$

$(s, p) = (3, 4)$

$\alpha = 2\cos(\pi/18)/\sqrt{3}$

$$\begin{array}{c|ccc} (1+\alpha)/2 & (1+\alpha)/2 & 0 & 0 \\ 1/2 & -\alpha/2 & (1+\alpha)/2 & 0 \\ (1-\alpha)/2 & 1+\alpha & -(1+2\alpha) & (1+\alpha)/2 \\ \hline & 1/(6\alpha^2) & 1-1/(3\alpha^2) & 1/(6\alpha^2) \end{array}. \tag{10.31}$$

定理的公式(10.31)说明，具有A稳定的四阶方法是由其三级公式给出来的.

定理 10.3 不存在具有 $(s, p)=(4, 5)$ 的 DIRK(4,5) 公式.

为了叙述下面的定理，假定A为s阶矩阵(a_{ij})，D为s阶对角矩阵 $\mathrm{diag}(c_1, \cdots, c_s)$，$b$为$s$维向量$(b_i)$，$e$为所有元素均为1的$s$维向量，$T$表示矩阵转置.

定理 10.4 令 $p \leqslant 5$，为了使 Runge-Kutta 方法(10.28)对于(10.23)中每个充分光滑的函数 $f(t, y)$ 有p阶导数，必须满足条件(10.32-1)—(10.32-5):

$$b^T e = 1, \tag{10.32-1}$$

$$b^T De = \frac{1}{2}, \quad b^T Ae = \frac{1}{2}; \tag{10.32-2}$$

$$\begin{aligned} &b^T D^2 e = 1/3, \ b^T DAe = 1/3, \\ &b^T ADe = 1/6, \ b^T A^2 e = 1/6; \end{aligned} \tag{10.32-3}$$

$$\begin{cases} b^T D^3 e = 1/4, \ b^T DADe = 1/8, \\ b^T AD^2 e = 1/12, \ b^T A^2 De = 1/24; \\ b^T D^2 Ae = 1/4, \ b^T DA^2 e = 1/8, \\ b^T ADAe = 1/12, \ b^T A^3 e = 1/24; \end{cases} \tag{10.32-4}$$

$$\begin{cases} b^T D^4 e = 1/5, \ b^T DAD^2 e = 1/15, \ b^T DA^2 De = 1/30; \\ b^T A^2 D^2 e = 1/60; \\ b^T D^3 Ae = 1/5, \ b^T DADAe = 1/15, \ b^T DA^3 e = 1/30; \\ b^T A^2 DAe = 1/60; \\ b^T D^2 ADe = 1/10, \ b^T AD^3 e = 1/20, \ b^T ADADe = 1/40; \\ b^T A^3 De = 1/120; \\ b^T D^2 A^2 e = 1/10, \ b^T AD^2 Ae = 1/20, \ b^T ADA^2 e = 1/40; \\ b^T A^4 e = 1/120. \end{cases} \tag{10.32-5}$$

在求解大的非线性刚性方程中，Prothero 和 Robinson[95] 发现A稳定的方法不保证给出稳定解，并且得到的解的精度也常常与

所用的方法的阶不相适应．由他们的工作导出了新的稳定性概念．

定义 10.2 Runge-Kutta 公式称为 S 稳定的，如果对任何具有有界导数的有界函数 $g:[0, T]\to R$ 和任何正的常数 λ_0，存在一个正常数 h_0，使方程

$$y' = \lambda(y - g(t)) + g'(t)$$

的数值解 $\{y_n\}$，对于 $y_n \neq g(t_n)$，以及所有 $0 < h < h_0$ 和所有满足 $\mathrm{Re}(-\lambda) \geqslant \lambda_0$ 的复数 λ 有

$$\left|\frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)}\right| < 1.$$

定义 10.3 Runge-Kutta 公式称为强 S 稳定的(strongs-stable)，如果当 $\mathrm{Re}(-\lambda)\to\infty$ 时，以及对于任何 $h > 0$ 和 $[t_n, t_{n+1}]\subset[0, T]$ (当 $n\to\infty$ 时)，有

$$\left|\frac{y_{n+1} - g(t_{n+1})}{y_n - g(t_n)}\right| \to 0.$$

注意，S 稳定的方法是 A 稳定的(取 $g \equiv 0$)，而反过来则不成立．

我们知道，对于每一个 Runge-Kutta 方法，取步长 h，将其应用到试验方程

$$y' = \lambda y,\ y(0) = y_0,$$

都可得到有理函数 $R(h\lambda)$:

$$R(h\lambda) = 1 + h\lambda b^T(I - h\lambda A)^{-1}e, \tag{10.33}$$

$$y_{n+1} = R(h\lambda)y_n.$$

正如前面所指出的，Runge-Kutta 公式称为 A 稳定的，如果 $\mathrm{Re}(h\lambda) < 0$ 时，有 $|R(h\lambda)| < 1$．Runge-Kutta 公式称为刚性 A 稳定的(或左稳定的，有的作者也称为强 A 稳定的)，如果它是 A 稳定的，且

$$\lim_{\substack{h\lambda\to\infty \\ \mathrm{Re}(h\lambda)<0}} R(h\lambda) = 0.$$

定义 10.4 具有非零对角线的半隐式公式，且 A 为可逆矩阵，它的 $R(h\lambda)$ 在无穷远点是正则的，且有

$$\alpha_0 = \lim_{h\lambda\to\infty} R(h\lambda) = 1 - b^T A^{-1} e. \tag{10.34}$$

当 $c_s=1$ 和 $a_{si}=b_i,\ i=1,\cdots,s$ 时，称 s 级半隐式 Runge-Kutta 方法是刚性精确的。

引理 10.1 具有满足 $a_{si}=b_i,\ i=1,\cdots,s$ 的可逆矩阵 A 的 Runge-Kutta 公式，有 $\alpha_0=0$. 特别，如果公式是 A 稳定的和刚性精确的，它是刚性 A 稳定的。

证明 结果立即可以得到. 因为由 (10.34)，$b^TA^{-1}=(0,0,\cdots,0,1)$，即 A 的最后一行乘 A^{-1}，给出单位矩阵的最后一行。

下面我们给出 Alexander[21] 关于判断稳定性的一些条件，而不加证明。

定理 10.5 具有正对角元素的 A 稳定的半隐式 Runge-Kutta 公式为 S 稳定的充要条件是 $|\alpha_0|<1$. 这一类的 S 稳定的公式是强 S 稳定的充要条件为它是刚性精确的。

现在我们把这些准则应用到 Crouzeix 所导出的 DIRK 公式上，得到如下的推论

推论 1 具有 $(s,p)=(2,3)$ 的 A 稳定的方法是 S 稳定的；具有 $(s,p)=(3,4)$ 的 A 稳定的方法是 S 稳定的，除了对于 $\alpha_0=1$ 的两个方法以外，这些方法中没有一个是强 S 稳定的。

定理 10.6 恰好存在两个二级二阶的强 S 稳定的 DIRK (2, 2) 公式和一个三级三阶的强 S 稳定的 DIRK (3, 3) 公式. 它们是

$$\begin{array}{c|cc} \alpha & \alpha & 0 \\ 1 & 1-\alpha & \alpha \\ \hline & 1-\alpha & \alpha \end{array}, \qquad \alpha = 1 \pm \frac{1}{2}\sqrt{2},$$

$$\begin{array}{c|ccc} \alpha & \alpha & 0 & 0 \\ c_2 & c_2-\alpha & \alpha & 0 \\ 1 & b_1 & b_2 & \alpha \\ \hline & b_1 & b_2 & \alpha \end{array}, \qquad \alpha \text{ 是 } x^3-3x^2+\frac{3}{2}x-\frac{1}{6}=0 \text{ 在 } \left(\frac{1}{6},\frac{1}{2}\right) \text{ 中的根},$$

$$c_2=(1+\alpha)/2,$$
$$b_1=-(6\alpha^2-16\alpha+1)/4,$$
$$b_2=(6\alpha^2-20\alpha+5)/4.$$

定理 10.7 不存在四级四阶的强 S 稳定的 DIRK 公式。

这个定理说明,要达到四阶的强 S 稳定的 DIRK 公式至少需要五级。

现在我们来叙述 DIRK 公式在计算机上的实现. 在开始时由使用者选取下面的 5 个公式中的一个.

1. DIRK (1, 2) 隐式中点法.
2. DIRK (2, 3) Crouzeix 公式(定理 10.2).
3. DIRK (3, 4) Crouzeix 公式(定理 10.2).
4. DIRK (2, 2) 定理 10.5 的强 S 稳定的公式,

$$\alpha = 1 \pm \frac{1}{2}\sqrt{2}.$$

5. DIRK (3, 3) 定理 10.5 的强 S 稳定的公式.

我们知道, DIRK (1, 2) 是 A 稳定的, 但不是 S 稳定的. 积分时按照缩小一倍步长的方法来估计误差和调整步长. 首先在时刻 t_n 处取步长 h, 由 y_n 计算 t_{n+1} 时刻的值记为 $y_{n+1}^{(h)}$. 其次, 再用步长 $h/2$ 积分两步由 y_n 计算 t_{n+1} 时刻的值记为 $y_{n+1}^{(h/2)}$. 更精确的 $y_{n+1}^{(h/2)}$ 的局部截断误差是

$$E_{n+1} = \|y_{n+1}^{(h)} - y_{n+1}^{(h/2)}\|/(2^p - 1),$$

其中 p 为方法的阶. $\|\cdot\|$ 是加权的均方根模

$$\|y\| = \left(\frac{1}{m}\sum_{i=1}^{m} (y^i/y_{\max})^2\right)^{1/2}.$$

$y_{\max}$ 是积分到当前 m 个分量中最大模的分量.

局部截断误差 E_{n+1} 用来调整步长, 使用者可以指定一个精度 ε.

(1) 如果 $E_{n+1} > \varepsilon$, 则这一步被舍弃, 然后将步长 h 缩小到产生所希望的误差 $\sim \varepsilon/5$.

(2) 如果 $3\varepsilon/4 < E_{n+1} \leqslant \varepsilon$, 则这一步就被接受, 但将步长缩小到使下一步所希望的误差 $\sim \varepsilon/5$.

(3) 如果 $\varepsilon/10 < E_{n+1} \leqslant 3\varepsilon/4$, 则这一步被采用, 并且在下一步应用同样的步长.

(4) 如果 $E_{n+1} \leqslant \varepsilon/10$，则这一步被接受，并将步长 h 放大，使得在下一步所希望的误差 $\sim \varepsilon/2$，还使得有

(i) 至少在 h 的最后一次减小后，有 $p+1$ 步成功.

(ii) 在减小后，在下一次放大 h 时，至多放大二倍，而在任何情形下，最大的放大为 10 倍.

(iii) 每次放大至少为因子 1.3.

每一级的隐式方程均用 Newton 迭代方法求解，矩阵

$$I - \alpha h \partial f/\partial y \quad 和 \quad I - \frac{1}{2}\alpha h \partial f/\partial y$$

由使用者提供的 Jacobi 矩阵来计算，它们的 LU 分解可以存贮起来重复使用，每 20 次更新一次或者在步长改变时更新. $y_{n,i}$ 的初始值可应用存贮的导数值的线性内插和外插的组合得到. 如果在三步 Newton 迭代后方法仍不收敛，则就应该更新 Jacobi 矩阵.

1977 年，Alexandex[95] 通过对许多种问题的试验指出:

(1) 对于高频振荡的问题，由于 DIRK 方法是 A 稳定的，所以它胜过 Gear 方法.

(2) 对于中等精度的问题，DIRK 方法可以与 Gear 方法相比较，但是对于高精度的问题，Gear 方法比较好.

(3) DIRK 方法要求的函数求值次数及调用 Jacobi 的次数较多，但是与 Gear 方法相比程序本身的工作量较少. 对于大的问题 DIRK 方法不如 Gear 方法有效.

§ 4 Rosenbrock 的半隐式 Runge-Kutta 方法

为了数值求解自守系统

$$y' = f(y), \quad y(0) = y_0 \tag{10.35}$$

其中 y, f 均是 m 维向量. 在 1964 年 Butcher 提出高阶隐式 Runge-Kutta 方法之前，在 1963 年 Rosenbrock[97] 提出了一个介于显式公式与隐式公式之间的 Runge-Kutta 公式. 1969 年 Haines[61]

进一步完善了这类公式．这类半隐式的 Runge-Kutta 公式既保持隐式 Runge-Kutta 公式的 A 稳定性性质，又避免了迭代．它是隐式 Runge-Kutta 公式实现的一个有效的方法，其一般形式为

$$y_{n+1} = y_n + \sum_{i=1}^{s} W_i K_i, \tag{10.36}$$

其中

$$\begin{aligned}
K_1 &= h[f(y_n) + b_1 J(y_n) K_1], \\
K_2 &= h[f(y_n + \beta_{21} K_1) + b_2 J(y_n + \eta_{21} K_1) K_2], \\
&\vdots \\
K_s &= h\left[f\left(y_n + \sum_{i=1}^{s-1} \beta_{si} K_i\right) + b_s J\left(y_n + \sum_{i=1}^{s-1} \eta_{si} K_i\right) K_s\right],
\end{aligned} \tag{10.37}$$

这里 $J = \partial f/\partial y$ 为 Jacobi 矩阵．

如果 (10.37) 中 $b_1 = b_2 = \cdots = b_s = 0$，则 (10.36)、(10.37) 成为显式 Runge-Kutta 公式．(10.37) 的每个 K_i 的公式中，在其右端也只包含未知量 $K_1, \cdots, K_i$，而不包含未知量 $K_{i+1}, K_{i+2}, \cdots, K_s$．所以，实际计算可将 $K_1, \cdots, K_s$ 从相应的公式中解出来．于是

$$\begin{aligned}
K_1 &= h[I - hb_1 J(y_n)]^{-1} f(y_n), \\
K_2 &= h[I - hb_2 J(y_n + \eta_{21} K_1)]^{-1} f(y_n + \beta_{21} K_1), \\
&\vdots \\
K_s &= h\left[I - hb_s J\left(y_n + \sum_{i=1}^{s-1} \eta_{si} K_i\right)\right]^{-1} f\left(y_n + \sum_{i=1}^{s-1} \beta_{si} K_i\right).
\end{aligned} \tag{10.38}$$

Rosenbrock 提出这类方法的动机是想得到稳定的 Runge-Kutta 方法，同时又要计算方便，即避免解隐式方程所需要的迭代．公式中的参数 $W_1, \cdots, W_s$；$b_1, \cdots, b_s$；$\beta_{21}, \beta_{31}, \cdots, \beta_{s1}, \cdots, \beta_{ss-1}$；$\eta_{21}, \eta_{31}, \eta_{32}, \cdots, \eta_{s1} \cdots, \eta_{ss-1}$，可以用与通常的 Runge-Kutta 方法跟 Taylor 级数进行比较的类似方法求出来．[71]给出了这类半隐式

Runge-Kutta 方法的几组公式，并分析了它们的稳定性，现摘要介绍如下：

Rosenbrock 二级三阶半隐式公式

$$b_1 = 1 + \sqrt{6}/6 = 1.40824829,$$

$$b_2 = 1 - \sqrt{6}/6 = 0.59175171,$$

$$\beta_{21} = \eta_{21} = \{-6 - \sqrt{6} + \sqrt{58 + 20\sqrt{6}}\}/(6 + 2\sqrt{6}) = 0.17378667,$$

$$\begin{aligned} W_1 &= -0.41315432, \\ W_2 &= 1.41315432. \end{aligned} \tag{10.39}$$

Rosenbrock 二级二阶半隐式公式

$$\begin{aligned} b_1 = b_2 &= 1 - \sqrt{2}/2, \\ \beta_{21} &= (\sqrt{2} - 1)/2, \\ \eta_{21} &= 0, \\ W_1 &= 0, \\ W_2 &= 1. \end{aligned} \tag{10.40}$$

Haines[61] 推导了类似的方法，并且包含了误差估计．他得到的一组参数如下：

Haines 四级三阶半隐式公式

$$\begin{aligned} & b_1 = 1,\ b_2 = 1,\ b_3 = 1,\ b_4 = \frac{2}{3}, \\ & \beta_{21} = 1, \\ & \beta_{31} = \frac{1}{2},\ \beta_{32} = \frac{1}{2}, \\ & \beta_{41} = \frac{2}{99},\ \beta_{42} = \frac{95}{99},\ \beta_{43} = \frac{2}{99}, \\ & \eta_{21} = 1, \\ & \eta_{31} = \frac{1}{2},\ \eta_{32} = \frac{1}{2}, \\ & \eta_{41} = 0,\ \eta_{42} = 0,\ \eta_{43} = 0, \end{aligned} \tag{10.41}$$

$$W_1=\frac{19}{9},\ W_2=-\frac{43}{18},\ W_3=\frac{28}{9},\ W_4=-\frac{11}{6}.$$

Calahan[40] 提出了一个二级三阶公式

$$b_1=b_2=\frac{1}{2}(1+\sqrt{1/3})=0.788675135,$$

$$\beta_{21}=-2\sqrt{1/3}=-1.154700538, \tag{10.42}$$

$$\eta_{21}=0,\ W_1=\frac{3}{4}=0.75,\ W_2=0.25.$$

将此类方法应用到试验方程 $y'=\lambda y$ 时，就导出形式如

$$y_{n+1}=R(h\lambda)y_n$$

的关系式，其中 $R(h\lambda)$ 为 $h\lambda$ 的有理多项式．如果 $\mathrm{Re}(h\lambda)<0$ 时有 $|R(h\lambda)|<1$，则方法是 A 稳定的．为了证明这一点，只要证明当 $\mu\to-\infty$ 时 $(\mu=h\lambda,\ h>0)$，$|R(\mu)|<1$.

对应于 (10.39) 的

$$R(h\lambda)=\frac{1-h\lambda-\frac{2}{3}h^2\lambda^2}{1-2h\lambda+\frac{5}{6}h^2\lambda^2},$$

显然，当 $\mathrm{Re}(h\lambda)<0$ 时，$|R(h\lambda)|<1$．所以，方法是 A 稳定的．

对应于 (10.40) 的

$$R(h\lambda)=\frac{1+(\sqrt{2}-1)h\lambda}{1+(\sqrt{2}-2)h\lambda+\left(\frac{3}{2}-\sqrt{2}\right)h^2\lambda^2},$$

这样，当 $\mathrm{Re}(h\lambda)<0$ 时，$|R(h\lambda)|<1$，而且可以看到当 $\mathrm{Re}(h\lambda)\to-\infty$ 时，$|R(h\lambda)|\to0$，所以方法还是 L 稳定的．上述两个方法的稳定区域，如图所示．

对于 Haines 方法，有

$$R(h\lambda)=\frac{1-\frac{8}{3}h\lambda+\frac{2}{9}h^2\lambda^2+\frac{1}{3}h^3\lambda^3}{1-\frac{11}{3}h\lambda+5h^2\lambda^2-3h^3\lambda^3+\frac{2}{3}h^4\lambda^4},$$

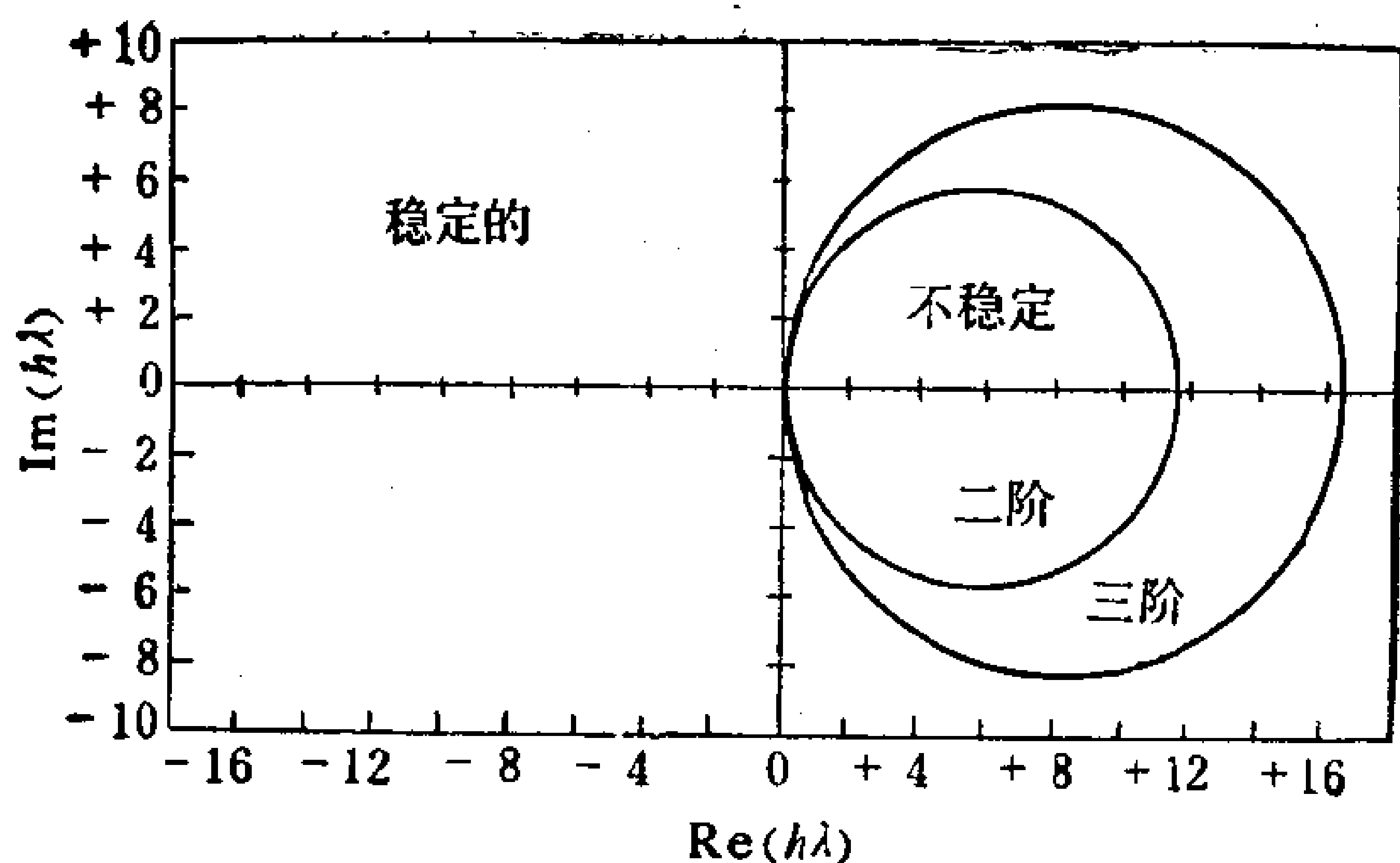


图 10.1 二阶和三阶 Rosenbrock 方法的稳定区域

方法仅在虚轴上有一个很小的不稳定区域，且是 L 稳定的．对于 Calahan 方法，有

$$R(h\lambda)=\frac{1-0.578h\lambda-0.456h^2\lambda^2}{1-1.578h\lambda+0.622h^2\lambda^2},$$

方法是 A 稳定的．

从上面所给出的公式中可以看出，这一类方法具有隐式 Runge-Kutta 方法的重要的稳定性性质，并且它避免了解非线性联立方程组．在实际的计算过程中，Jacobi 矩阵 $J=\partial f/\partial y$ 不需要每次都重新计算．特别对于常系数线性微分方程组 $y'=Ay$，相应的逆矩阵 A^{-1} 只需要计算一次．

§5 Butcher 矩阵变换及相应的方法

不失一般性，我们考虑 m 维自守系统的初值问题

$$y'=f(y),\ y(0)=y_0. \tag{10.43}$$

由 s 级隐式 Runge-Kutta 方法

$$K_i=y_{n-1}+h\sum_{j=1}^{s}a_{ij}f(K_j),\ i=1,\cdots,s, \tag{10.44}$$

$$y_n=y_{n-1}+h\sum_{j=1}^{s}b_jf(K_j) \tag{10.45}$$

得到(10.43)的近似解．为了求出满足(10.44)的 K_j，$j=1,\cdots,s$，通常采用 Newton-Raphson 迭代方法．假设经过一次迭代后 K_j 的修正值为 $K_j+\delta_j$，$j=1,\cdots,s$，而 δ_j 由下式确定

$$\begin{bmatrix} I-ha_{11}J_1 & -ha_{12}J_2 & \cdots & -ha_{1s}J_s \\ -ha_{21}J_1 & I-ha_{22}J_2 & \cdots & -ha_{2s}J_s \\ \vdots & & & \\ -ha_{s1}J_1 & -ha_{s2}J_2 & \cdots & I-ha_{ss}J_s \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_s \end{bmatrix}$$

$$+\begin{bmatrix} K_1-y_{n-1}-h\sum_{j=1}^{s}a_{1j}f(K_j) \\ \vdots \\ K_s-y_{n-1}-h\sum_{j=1}^{s}a_{sj}f(K_j) \end{bmatrix}=0, \qquad (10.46)$$

其中

$$J_i=\begin{bmatrix} \frac{\partial f_1}{\partial y_1}(K_i) & \cdots & \frac{\partial f_1}{\partial y_m}(K_i) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial y_1}(K_i) & \cdots & \frac{\partial f_m}{\partial y_m}(K_i) \end{bmatrix} \text{为}\, m\, \text{阶矩阵}, \ i=1,\cdots,s.$$

为了下面叙述方便(10.46)式可写成缩简的形式

$$\delta_i-h\sum_{j=1}^{s}a_{ij}J_j\delta_j-z_i=0, \qquad (10.47)$$

这里

$$z_i=-K_i+y_{n-1}+h\sum_{j=1}^{s}a_{ij}f(K_j), \ i=1,\cdots,s. \qquad (10.48)$$

使用 Newton-Raphson 方法求解 ms 阶非线性方程组，计算时间主要花在 Jacobi 矩阵的求值和线性代数方程组(10.47)的处理上．1976年，Butcher[35] 提出了一种矩阵变换方法，即对(10.47)预先做适当的变换，可以使迭代所花的工作量大大减少．下面简要摘录[35]中介绍的有关的一些结果．

假设 s 阶矩阵为 $A=(a_{ij})$，ms 维向量 δ，z 为

$$\delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_s \end{pmatrix}, \quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_s \end{pmatrix},$$

则 (10.47) 式为

$$\begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_s \end{pmatrix} - h \begin{pmatrix} a_{11}J_1 & a_{12}J_2 & \cdots & a_{1s}J_s \\ a_{21}J_1 & a_{22}J_2 & \cdots & a_{2s}J_s \\ \vdots & & & \vdots \\ a_{s1}J_1 & a_{s2}J_2 & \cdots & a_{ss}J_s \end{pmatrix} \begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_s \end{pmatrix} - \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_s \end{pmatrix} = 0, \quad (10.49)$$

这儿 δ_i，z_i 是 m 维向量，J_i 是 m 阶矩阵，$i = 1, \cdots, s$. 若只考虑 $J = J_1 = J_2 = \cdots = J_s$ 的情形，则有

$$\begin{pmatrix} a_{11}J_1 & \cdots & a_{1s}J_s \\ \vdots & & \vdots \\ a_{s1}J_1 & \cdots & a_{ss}J_s \end{pmatrix} = \begin{pmatrix} a_{11}J & \cdots & a_{1s}J \\ \vdots & & \\ a_{s1}J & \cdots & a_{ss}J \end{pmatrix} = A \otimes J,$$

(10.49) 式变为

$$\delta - hA \otimes J\delta - z = 0.$$

若令

$$M = \tilde{I} \otimes I - hA \otimes J,$$

则方程组 (10.47) 可写成

$$M\delta - z = 0, \quad (10.50)$$

这里 $\tilde{I}$ 是 s 阶单位矩阵，I 是 m 阶单位矩阵.

以下假设 A 是非奇异的，P、Q 是待定的 s 阶矩阵，又令

$$\tilde{\delta} = (Q^{-1} \otimes I)\delta, \quad \tilde{z} = (P \otimes I)z,$$
$$\widetilde{M} = (P \otimes I)M(Q \otimes I),$$

把 M 代入上式，有

$$\begin{aligned} \widetilde{M} &= (P \otimes I)(\tilde{I} \otimes I - hA \otimes J)(Q \otimes I) \\ &= (P\tilde{I} \otimes I - hPA \otimes J)(Q \otimes I) \\ &= (PQ) \otimes I - h(PAQ) \otimes J. \end{aligned}$$

再令 $\widetilde{A} = PAQ$，则

$$\widetilde{M} = (PQ) \otimes I - h\widetilde{A} \otimes J.$$

于是，方程组 (10.50) 变成为

$$\widetilde{M}\tilde{\delta} - \tilde{z} = 0, \tag{10.51}$$

其中 z 是已知量.

下面我们分析如何选取矩阵 P 和 Q 来减少计算量. 设 A, R 是非奇异的, A^{-1} 的广义的 Jordan 标准型为

$$R^{-1}A^{-1}R = \begin{bmatrix} \lambda_1^{-1} & & & \mathbf{0} \\ \mu_1 & \lambda_2^{-1} & & \\ & \ddots & \ddots & \\ \mathbf{0} & & \mu_{s-1} & \lambda_s^{-1} \end{bmatrix},$$

这里 λ_i 是 A 的特征值. 当 $\lambda_i \neq \lambda_{i+1}$ 时, $\mu_i = 0$, $i = 1, \cdots, s$, 当 $\lambda_i = \lambda_{i+1}$ 时, μ_i 为零或为任意的非零数, 这里 μ_i 非零时取 $\mu_i = \lambda_i^{-1}$.

令 $D = \operatorname{diag}(\lambda_1, \cdots, \lambda_s)$, 并选取 $P = DR^{-1}A^{-1}$, $Q = R$, 则有

$$PQ = \begin{bmatrix} 1 & & & \mathbf{0} \\ \varepsilon_1 & 1 & & \\ & \ddots & \ddots & \\ \mathbf{0} & & \varepsilon_{s-1} & 1 \end{bmatrix},$$

这里下对角线元素 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_{s-1}$ 是 0 或者 1,且有

$$PAQ = D.$$

这样, (10.51) 中矩阵 $\widetilde{M}$ 的对角元素为形如 $I - h\lambda J$ 的子块,下对角元素为零块或 I, 其余都是零块. 在这种情形下,就变得只要对每个对角块进行 LU 分解和分割成 s 个独立的子块的回代过程. 显然,这比直接对 ms 阶矩阵 M 进行 LU 分解要节省许多计算量.

从上述的过程中看到, 为了节省计算机时我们希望矩阵 A 只有实特征值, 特别地恰好只有一个 s 重的实特征值的情形是很适合于 Butcher 提出的这种矩阵变换. Burrage[37] 提出的简单隐式方法和 Nørsett[92] 提出的配置方法就是具有这种性质的 Runge-Kutta 方法. 下面我们仅介绍简单隐式 Runge-Kutta 方法.

1978 年 Burrage[37] 提出了简单隐式 Runge-Kutta 方法. 它是隐式 Runge-Kutta 方法实现的比较有效的方法. 在这种方法类

中，虽然矩阵A没有半隐式 Runge-Kutta 方法的三角形结构，但是它具有只有一个s重根的特征多项式．由于这种方法结构简单，容易实现变阶变步长的积分程序包．

为了构造这种方法类，现做如下的简化假定：

定义 10.5 以 $C(p)$，$D(p)$，$B(p)$ 表示以下的条件，

$C(p)$：$\sum_{j=1}^{s} a_{ij}c_j^{k-1}=c_i^k/k$，对于 $i=1,\cdots,s$ 和 $k=1,2,\cdots,p$.

$D(p)$：$\sum_{i=1}^{s} b_ic_i^{k-1}a_{ij}=b_j(1-c_j^k)/k$，对于 $j=1,\cdots,s$，和 $k=1,2,\cdots,P$.

$B(p)$：$\sum_{i=1}^{s} b_ic_i^{k-1}=1/k$，对于 $k=1,2,\cdots,P$.

为了以下讨论方便，我们不加证明地引用 Butcher[33] 最早发表的关于s级 Runge-Kutta 方法的如下结果：

引理 10.2 命题 $C(\eta)$，$D(\xi)$，$B(p)$ 成立，其中 $p\leqslant\xi+\eta+1$，$p\leqslant 2\eta+2$ 表示方法具有p阶．

从这个引理可以看到，对于 $t=0,\cdots,s$，则由条件 $C(s-1)$，$D(t)$，$B(s+t)$ 表示方法具有 $s+t$ 阶．如果节点 $c_1,\cdots,c_s$ 是互异的，在定理 10.8 中将看到，A必将是一个类似于特殊类型的矩阵．命题 $C(s-1)$，$D(t)$，$B(s+t)$成立，且 $c_1,\cdots,c_s$ 互异的方法称为转换方法．

定义 10.6 若s阶的或更高阶的转换方法，其 Runge-Kutta 矩阵正好有一个实的s重的特征值，则这个方法称为简单隐式方法．

定理 10.8 s级s阶的或更高阶的转换方法族由

$$\begin{array}{c|c} c_1 & \\ \vdots & V_sA_sV_s^{-1} \\ c_s & \\ \hline & b_1\ \cdots\ b_s \end{array}$$

给出，其中 $(b_1,\cdots,b_s)=(1,\cdots,1/s)V_s^{-1}$，

$$V_s=\begin{bmatrix}1 & c_1\cdots c_1^{s-1}\\ 1 & c_2\cdots c_2^{s-1}\\ & \cdots \\ 1 & c_s\cdots c_s^{s-1}\end{bmatrix},\quad A_s=\begin{bmatrix}0 & 0 & \cdots & 0 & \alpha_{1s}\\ 1 & 0 & & \vdots & \vdots\\ 0 & 1/2 & & \vdots & \vdots\\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & 1/(s-1) & \alpha_{ss}\end{bmatrix},$$

$\alpha_{js}\in R,\ j=1,\cdots,s$.

证明　由转换方法的节点 $c_1,\cdots,c_s$ 是互异的，所以 A 可写成

$$A=V_sA_sV_s^{-1}, \tag{10.52}$$

其中 V_s 是 Vandermonde 矩阵，A_s 的元素以 α_{ij} 表示。

令 $e_1,e_2,\cdots,e_s$ 是标准的列基向量，又令 $C^k=(c_1^k,\cdots,c_s^k)^T$，$k=0,1,\cdots$. 因此条件 $C(s-1)$ 与

$$AC^{k-1}=C^k/k,\ k=1,\cdots,s-1 \tag{10.53}$$

是等价的. 由 (10.52)

$$A_sV_s^{-1}C^{k-1}=\frac{1}{k}V_s^{-1}C^k,$$

但

$$V_s^{-1}C^{k-1}=e_k,\ V_s^{-1}C^k=e_{k+1}.$$

故有

$$A_se_k=\frac{1}{k}e_{k+1},\ k=1,\cdots,s-1.$$

由此可得 A_s 前 $s-1$ 列的元素，

$$A_s=\begin{bmatrix}0 & 0 & \cdots & 0 & \alpha_{1s}\\ 1 & 0 & & & \vdots\\ 0 & 1/2 & & & \vdots\\ \vdots & \vdots & \ddots & & \vdots\\ 0 & 0 & \cdots & 1/(s-1) & \alpha_{ss}\end{bmatrix},\ \alpha_{is}\in R,\ i=1,\cdots,s. \tag{10.54}$$

由定义 10.5，条件 $B(s+t)$ 是

$$\sum_{i=1}^{s}b_ic_i^{k-1}=\frac{1}{k},\ k=1,\cdots,s+t.$$

因为节点 c_i 是互异的，前 s 个等式是

$$(b_1, \cdots, b_s)V_s = \left(1, \cdots, \frac{1}{s}\right)$$

或

$$(b_1, \cdots, b_s) = (1, \cdots, 1/s)V_s^{-1}, \tag{10.55}$$

最后，条件 $D(t)$ 就是

$$k(b_1c_1^{k-1}, \cdots, b_sc_s^{k-1})A = (b_1 - b_1c_1^k, \cdots, b_s - b_sc_s^k),$$
$$k = 1, \cdots, t.$$

把 $k = 1, 2, \cdots, t$ 写在一起，即

$$\begin{pmatrix} 1\cdot b_1 & \cdots & 1\cdot b_s \\ 2b_1c_1 & \cdots & 2b_sc_s \\ \cdots & \cdots & \cdots \\ tb_1c_1^{t-1} & \cdots & tb_sc_s^{t-1} \end{pmatrix} A = \begin{pmatrix} b_1(1-c_1) & \cdots & b_s(1-c_s) \\ b_1(1-c_1^2) & \cdots & b_s(1-c_s^2) \\ \cdots & \cdots & \cdots \\ b_1(1-c_1^t) & & b_s(1-c_s^t) \end{pmatrix},$$

将 A 代成 $V_sA_sV_s^{-1}$，再右乘 V_s，则得

$$\begin{pmatrix} 1\cdot\Sigma b_i & 1\cdot\Sigma b_ic_i & \cdots & 1\cdot\Sigma b_ic_i^{s-1} \\ 2\Sigma b_ic_i & 2\Sigma b_ic_i^2 & \cdots & 2\Sigma b_ic_i^s \\ t\Sigma b_ic_i^{t-1} & t\Sigma b_ic_i^t & \cdots & t\Sigma b_ic_i^{s+t-2} \end{pmatrix} A_s$$
$$= \begin{pmatrix} \Sigma b_i(1-c_i) & \Sigma b_i(1-c_i)c_i & \cdots & \Sigma b_i(1-c_i)c_i^{s-1} \\ \Sigma b_i(1-c_i^2) & \Sigma b_i(1-c_i^2)c_i & \cdots & \Sigma b_i(1-c_i^2)c_i^{s-1} \\ \Sigma b_i(1-c_i^t) & \Sigma b_i(1-c_i^t)c_i & \cdots & \Sigma b_i(1-c_i^t)c_i^{s-1} \end{pmatrix},$$

应用条件 $B(p)$，$k = 1, \cdots, s+t$，有

$$\begin{pmatrix} 1\cdot 1 & 1\cdot\frac{1}{2} & \cdots & 1\cdot\frac{1}{s} \\ 2\cdot\frac{1}{2} & 2\cdot\frac{1}{3} & \cdots & 2\cdot\frac{1}{s+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ t\cdot\frac{1}{t} & t\cdot\frac{1}{t+1} & \cdots & t\cdot\frac{1}{s+t-1} \end{pmatrix} A_s$$

$$
=\begin{pmatrix}
1-\frac{1}{2} & \frac{1}{2}-\frac{1}{3} & \cdots & \frac{1}{s}-\frac{1}{s+1} \\
1-\frac{1}{3} & \frac{1}{2}-\frac{1}{4} & \cdots & \frac{1}{s}-\frac{1}{s+2} \\
\cdots & \cdots & \cdots & \cdots \\
1-\frac{1}{t+1} & \frac{1}{2}-\frac{1}{t+2} & \cdots & \frac{1}{s}-\frac{1}{s+t}
\end{pmatrix}.
$$

写出与 A_s 最后一列有关的 t 个等式,即

$$
\sum_{j=1}^{s} \frac{\alpha_{js}}{k+j-1}=\frac{1}{s(s+k)}, \quad k=1,\cdots,t. \qquad (10.56)
$$

定理证毕.

定理 10.9 s 阶的或更高阶的简单隐式方法类由定理 10.8 给出,并有

$$
\alpha_{ks}=(-1)^{s-k}\binom{s}{k-1}\frac{(s-1)!}{(k-1)!}\lambda^{s-k+1}, \quad k=1,\cdots,s,
$$

$\lambda\in R-\{0\}$.

证明 根据定义 10.6,简单隐式方法有互异的节点 c_i. 因此, $A=V_sA_sV_s^{-1}$ 是相似变换, 所以 A 的特征多项式和 A_s 的特征多项式相同. 但 A_s 的特征多项式是

$$
\rho(z)=z^s-\sum_{k=1}^{s}\alpha_{ks}(k-1)!\,z^{k-1}/(s-1)!.
$$

简单隐式方法应有

$$
\rho(z)=(z-\lambda)^s.
$$

比较 z 的系数便得到定理的结果.

定义 10.7 次数为 s 的多项式 $L_s(z)$, 定义为

$$
L_s(z)=\sum_{j=0}^{s}(-1)^j\binom{s}{j}z^j/j!,
$$

对于整数 n 和 m, $0\leqslant m\leqslant n$, 我们还定义多项式为

$$
\rho_n^{(m)}(z)=\sum_{j=0}^{n-m}(-1)^j\binom{n-m}{j}z^j/(n-j)!.
$$

由这个定义，我们知道如果 $L_n^{(m)}(z)$ 表示 $L_n(z)$ 的第m次导数，那么

$$\rho_n^{(m)}(z) = (-1)^n(n-m)!\,z^{n-m}L_n^{(m)}\left(\frac{1}{z}\right)\Big/ n!$$

推论 简单隐式方法，如果它具有定理 10.9 给出的形式，且条件 $D(1)$，$B(s+1)$ 成立，则方法为 $s+1$ 阶.

证明 由条件 $D(1)$ 与 (10.56)，$t=1$，则有

$$\sum_{k=1}^{s}\frac{\alpha_{ks}}{k}-\frac{1}{s(s+1)}=0.$$

将定理 10.9 中的 α_{ks} 代入上式，得

$$\sum_{k=1}^{s}(-1)^{s-k}\frac{(s-1)!}{k!}\binom{s}{k-1}\lambda^{s-k+1}-\frac{1}{s(s+1)}=0$$

即

$$\sum_{k=0}^{s}(-1)^{s-1-k}\frac{(s-1)!}{(k+1)!}\binom{s}{k}\lambda^{s-k}=0$$

或者

$$\rho_{s+1}^{(1)}(\lambda)=0.$$

Burrage[37] 还证明了如下的定理，这里摘录如下，不加证明.

定理 10.10 简单隐式方法的最大阶数为 $s+1$.

现在我们来讨论简单隐式方法的稳定性. 对于给定的 λ，简单隐式方法的特点之一是，它们像半隐式 Runge-Kutta 一样，有同样形式的有理函数

$$R(z)=1+zb^T(I-zA)^{-1}e.$$

Nørsett[92] 研究的半隐式 Runge-Kutta 方法的A稳定性所取得的结果稍加修改就可应用于简单隐式方法.

定理 10.11 对应于 s 阶的简单隐式方法的有理函数，由

$$R(z)=(-1)^s\sum_{k=0}^{s}\lambda^k L_s^{(s-k)}(1/\lambda)z^k/(1-\lambda z)^s \quad (10.57)$$

给出.

按照这个定理的公式 (10.57)，若 $L_{s+1}^{(1)}(1/\lambda)=0$，则 $R(z)$

是 $s+1$ 阶的简单隐式方法的有理函数.

定理 10.12 s 阶的或更高阶的简单隐式方法为 A 稳定的充要条件是

1) 特征根 $\lambda>0$;

2) $|L_s(1/\lambda)|\leqslant 1$;

3) $(1-L_s(1/\lambda)^2)(\lambda_r)^{2s}+2\sum_{i=k}^{s-1}(-1)^{s+i-1}(\lambda r)^{2i}\int_0^{1/\lambda}L_s(y)\cdot L_s^{(2s+1-2i)}(y)dy\geqslant 0$ 对于所有的 $r\in R$, 其中 $k=[(s+2)/2]$.

Burrage[37] 指出,通过计算给出如下的 A 稳定的方法的 λ 值的范围.

表 10.2

s	λ
1	[1/2, ∞]
2	[1/4, ∞]
3	[1/3, 1.06858]
4	[0.39434, 1.28057]
5	[0.24651, 0.36180]∪[0.42079, 0.47328]
6	[0.28407, 0.54090]

我们注意,一般不要选取区域的边界上的 λ 值,因为 λ 的小的变化能引起非 A 稳定性. 在简单隐式方法中选取 λ 时,有两种自然的选取,即使得这个方法有 $s+1$ 阶 A 稳定的或者有 s 阶 L 稳定的. 因此,我们有如下的 λ 的数值表

表 10.3

s 级简单隐式方法	A 稳定的和 $p=s+1$	L 稳定的和 $p=s$
1	1/2	1
2	$(3+\sqrt{3})/6$	$(2\pm\sqrt{2})/2$
3	1.06858	0.43587
4	—	0.57282
5	0.47328	0.27805
6	—	0.33414

§6 广义 Runge-Kutta 方法

1967 年，Lawson[72] 利用变换的方法将刚性方程换成非刚性方程的思想构造了适合于解刚性方程的 A 稳定的广义 Runge-Kutta 方法。

考虑求解初值问题

$$y' = f(t, y),\ y(t_0) = y_0, \tag{10.58}$$

其中 y 和 f 是 m 维向量. 我们定义新变量

$$x(t) = \exp(-tA)y(t). \tag{10.59}$$

于是，(10.58) 变成为

$$x'(t) = \exp(-tA)[f(t, \exp(tA)x(t)) - A\exp(tA)x(t)], \tag{10.60}$$
$$x(t_0) = \exp(-t_0A)y_0.$$

这里 A 是待定的 m 阶矩阵。我们如果记 (10.60) 为 $x' = g(t, x)$，则其右函数的 Jacobi 矩阵为

$$g_x = \exp(-tA)[\partial f/\partial y - A]\exp(tA) \tag{10.61}$$

因此，g_x 的特征值为 $\partial f/\partial y - A$ 的特征值. 若选取 A 使得这些特征值都很小，则方程组 (10.60) 就是非刚性的.

将一般的显式 Runge-Kutta 公式

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_iK_i,$$

$$K_i = f\left(t_n + c_ih,\ y_n + h\sum_{j=1}^{i-1} a_{ij}K_j\right), \tag{10.62}$$

$$i = 1, \cdots, s,$$

$$0 = c_1 \leqslant c_2 \leqslant \cdots \leqslant c_s = 1$$

应用到 (10.60)，并将公式中关于 $\exp(tA)x(t)$ 的各项分别用 p_i 表示，我们得到

$$K_1 = \exp(-t_nA)[f(t_n, \exp(t_nA)x_n) - A\exp(t_nA)x_n],$$

$$p_i = \exp[(t_n + c_ih)A]\left[x_n + h\sum_{j=1}^{i-1} a_{ij}K_j\right], \tag{10.63}$$

$$K_i = \exp[-(t_n + c_i h)A][f(t_n + c_i h, p_i) - Ap_i],$$
$$i = 2, \cdots, s,$$
$$x_{n+1} = x_n + h\sum_{i=1}^{s} b_i K_i.$$

若将 K_i 中除去 $\exp[-(t_n + c_i h)A]$ 的因数表为 K_i^*，相当于 p_i 的量表为 $p_i^*(i = 1, \cdots, s)$，指数只用增量，并取某种近似. 这个算法也可以改写成更方便的形式

$$\begin{aligned}
K_1^* &= f(t_n, y_n) - Ay_n, \\
p_i^* &= E(c_i hA)y_n + h\sum_{j=1}^{i-1} a_{ij}E[(c_i - c_j)hA]K_j^*, \\
K_i^* &= f(t_n + c_i h, p_i^*) - Ap_i^*, \\
y_{n+1} &= E(hA)y_n + h\sum_{i=1}^{s} b_i E[(1 - c_i)hA]K_i^*,
\end{aligned} \tag{10.64}$$

其中 $E(hA)$ 是 $\exp(hA)$ 的近似. 矩阵 A 的最自然的选取是 f 的 Jacobi 矩阵 $\partial f/\partial y$. 但是，我们只需要 $\partial f/\partial y - A$ 的特征值小就行，所以这种选取并不要求很精确，可以积分一段时间计算一次. Lawson 证明，如果对于所有的 h，$\|E(hA)\| < 1$，则算法 (10.64) 是 A 稳定的. 指数近似 $E(hA)$ 最好取 Padé 近似.

由于方法中包含了 Padé 近似以及积分一段时间计算一次 Jacobi 矩阵. 因而具有 Rosenbrock 方法的类似性质. 但是 (10.64) 式出现在 p_i^* 和 y_{n+1} 中的求和可以看成是形式为

$$h\int_0^{\alpha} \exp[(\alpha - \tau)hA]K(t_n + \tau h)d\tau$$

的积分近似.

算法 (10.64) 的过程虽然是稳定的，但可能需要很小的积分步长，因此可能花费更多的计算机时. 为了克服这种缺点，1975 年，Ehle 和 Lawson[53] 提出将 (10.64) 式的最后一步用

$$y_{n+1} = E(hA)y_n + h\sum_{i=1}^{s} W_i(hA)K_i^* \tag{10.65}$$

来代替. 这里的矩阵 $W_i(hA)$ 具有如下的性质:

1) $W_i(0)=b_iI$，I 是单位矩阵；

2) $b_jW_i(hA)=b_iW_j(hA)$，如果 $c_i=c_j$；

3) $\sum_{i=1}^{s}W_i(hA)(c_i)^k=M_k$，$k=0,1,\cdots,v-1$，$v\leqslant s$；

其中 v 是方法中不同的 c_i 的个数。

而

$$M_0=(hA)^{-1}(E(hA)-I),$$

$$M_{i+1}=(hA)^{-1}((i+1)M_i-I),\quad i=0,1,\cdots,v-2.$$

这里 M_0 由 $\exp(hA)$ 的近似 $E(hA)$ 来确定，M_k 可逆推得到。由 2) 和 3) 式结合可以推出存在唯一的解 $\{W_i\}$，$i=1,\cdots,m$。

类似地也可将 $h\sum_{j=1}^{i-1}a_{ij}E[(c_i-c_j)hA]K_j^*$ 看成为 $h\int_0^{c_i}\exp[(c_i-\tau)hA]K^*(t_n+\tau h)d\tau$ 的近似，并且将其换成形式为 $h\sum_{j=1}^{i-1}V_{ij}(hA)K_j^*$ 的求积公式，$V_{ij}(hA)$ 与上面的 $W_i(hA)$ 类似地构造，即多项式 K_j^* 被精确地积分到与原始 Runge-Kutta 方法相同的阶。

例如，我们来考虑经典的四阶显式 Runge-Kutta 方法。由 $\exp(z)$ 的 Padé 近似

$$E_{3,1}(z)=[1-3z/4+z^2/4-z^3/24]^{-1}(1+z/4)$$

表示成为 $d_{3,1}^{-1}(z)n_{3,1}(z)$，我们得到

$$W_1(z)=d_{3,1}^{-1}(z)(1/6+z/24),$$

$$W_2(z)=W_3(z)=d_{3,1}^{-1}(z)(1/3-z/12),$$

$$W_4(z)=d_{3,1}^{-1}(z)(1/6-z/8+z^2/24).$$

在这种情形下，函数 $V_1(z)=d_{3,1}^{-1}(z)(1-z/4+z^2/24)$。这时广义 Runge-Kutta 公式的形式为

$$K_1^*=f(t_n,y_n)-Ay_n,$$

$$p_2^*=E_{3,1}\left(\frac{hA}{2}\right)y_n+\frac{h}{2}V_1\left(\frac{hA}{2}\right)K_1^*,$$

$$K_2^* = f\left(t_n + \frac{h}{2},\, p_2^*\right) - Ap_2^*,$$

$$p_3^* = E_{3,1}\left(\frac{hA}{2}\right)y_n + \frac{h}{2} V_1\left(\frac{hA}{2}\right)K_2^*,$$

$$K_3^* = f\left(t_n + \frac{h}{2},\, p_3^*\right) - Ap_3^*,$$

$$p_4^* = E_{3,1}(hA)y_n + hV_1(hA)K_3^*,$$

$$K_4^* = f(t_n + h,\, p_4^*) - Ap_4^*,$$

$$y_{n+1} = E_{3,1}(hA)y_n + h[W_1K_1^* + W_2(K_2^* + K_3^*) + W_4K_4^*].$$

Ehle 和 Lawson[53] 指出，对于线性问题，刚性比从 100 到 10000 时，广义 Runge-Kutta 方法与 Gear 方法两者差不多，但是当特征值虚部比较大时，这里介绍的广义 Runge-Kutta 方法比 Gear 方法要好。

本 章 附 注

§1 的材料主要取自 Chipman 的 [48]；

§2 的材料主要取自 Bickart 的 [28]；

§3 的材料主要取自 Alexander 的 [21]；

§4 的材料选自 Lapidus 等人的书 [71]；

§5 的材料主要取自 Butcher 的 [35]，Burrage 的 [37] 和汤怀民的 [12]；

§6 的材料主要取自 Lawson 的 [72] 和 Ehle、Lawson 的 [53]。

第十一章 组合方法

§1 例 子

许多系统的运动可以用一个常微分方程组的初值问题

$$x' = H(x, t), \quad x(t_0) = x_0, \quad x \in R^{m+n} \tag{11.1}$$

来描述．若系统很大，则微分方程的维数很高．常有这样的现象，有的变化很快，而有的变化很慢．假设方程组(11.1)可以分解成下面的形式的两个子系统

$$y' = f(y, z, t), \quad y(t_0) = y_0, \quad y \in R^m, \tag{11.2}$$

$$z' = g(y, z, t), \quad z(t_0) = z_0、 \quad z \in R^n, \tag{11.3}$$

其中(11.2)可以看成是刚性方程组，(11.3)可以看成是非刚性方程组．这种情形在自动控制系统数字仿真、电子网络、电路分析中都可看到．

例 11.1 飞行器的仿真

近代的飞行器是一个复杂的自动控制系统．它由许多不同的变化速率的子系统组成．一般，可以分为控制和受控对象两个子系统来考虑，如图 11.1 所示．控制子系统，如飞行稳定性控制，飞行轨道控制等一般是电子元件组成的，用电信号控制受控部件的运动，反应快，变化非常迅速． 描述这些状态变化的方程属于(11.2)，是刚性的．受控部件是飞行器本身，它的运动惯性较大，反应慢，变化比电信号慢得多． 描述这种状态变化的方程属于

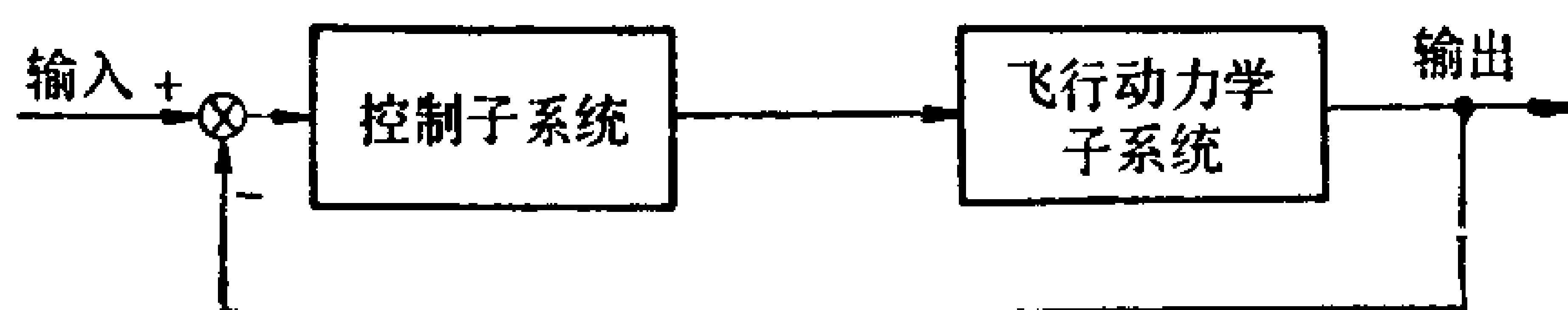


图 11.1 飞行器仿真分为两个子系统的框图

(11.3)，是普通的微分方程组，不是刚性的。

有些情形，也可以把系统分成更多个子系统来考虑。例如下面的例 11.2 就可以分成三个子系统来考虑。

例 11.2　设有一个六维系统

$$y' = A(y - \phi(t)) + \Phi'(t),\ y(0) = \Phi(0), \tag{11.4}$$

其中 A 为 6×6 阶常数矩阵，其特征值的实部为非正的，该系统的解 $\Phi(t)$ 为

$$\Phi(t) = [e^{-500t}\cos 5t,\ e^{-500t}\sin 5t,\ e^{-11t}\cos 2t,\ e^{-11t}\sin 2t,\ e^{-0.05t},\ e^{-0.01t}]^T.$$

解 $\Phi(t)$ 的分量中，前两个衰减得很快，中间两个以中等速度衰减，而后两个衰减得很慢。这样的系统，有人也称它为多种变化速率的系统。

由于上述子系统 (11.2) 是刚性的，因而整个系统 (11.1) 也是刚性的。所以需要采用适合于求解刚性方程的方法来求解。但是在很多系统中，刚性方程组 (11.2) 仅占整个方程组 (11.1) 的很小一部分，而且右函数相当简单。因而整个右函数的计算量主要集中在非刚性方程组 (11.3) 上。另一方面，(11.2) 右函数 Jacobi 矩阵特征值的绝对值要比 (11.3) 的大得多。整个方程组 (11.1) 的积分步长 h 的上界又要由 (11.2) 决定，需要取很小的步长 h，因而需要很大的计算量。所以对整个方程组采用同一个数值积分方法来处理是不合理的。1963 年在 [18] 中曾将 (11.1) 分解成子系统 (11.2) 和 (11.3)，并且分别对它们采用不同的数值方法来处理，即对 (11.2) 采用小步长的 Runge-Kutta 方法积分，而对 (11.3) 采用大步长的 Adams 方法积分，其中右函数计算中用到的另外一个子系统的值通过插值得到。[9] 中把这种对于不同的子系统采用不同的数值积分公式的方法，称为组合方法。Gear[57] 对具有不同变化速率的分量的系统也采用这种组合方法积分。

虽然目前已经有了许多适合于求解刚性方程的数值方法。但是，上述将方程分类，按不同的方法处理的思想对于相当多的问题是适宜的。理由是目前已有的刚性方程的求解方法，比起传统的

解非刚性方程的方法仍然要增加很多处理和计算量．特别是适合于求解刚性方程的方法一般都是隐式的，需要求解非线性方程组．由于迭代收敛性的要求，一般也不能使用简单迭代法，而要用 Newton-Raphson 方法，这就要求计算方程组(11.1)的右函数的 Jacobi 矩阵及相应的逆矩阵，其计算量是按方程组的维数的平方增加的．因此缩小刚性方程组的维数对于节省计算时间仍然有明显的实际的意义．本章主要讨论这方面的一些处理方法．

§2 基本算法公式

下面我们首先来描述用两个不同的数值方法分别积分子系统(11.2)和(11.3)的计算过程，即用刚性方法积分快变子系统(11.2)，用非刚性方法积分慢变子系统(11.3)的计算过程．假设我们选取数值解子系统(11.2)的刚性方法为 F，数值解子系统(11.3)的非刚性方法为 S(F，S 分别取 Fast 和 Slow 的字头，为快变、慢变之意)．它们用的积分步长分别取为 h 和 H，$H=rh$，r 为取定的整数．假定 $t_{k,F}$ 是方法 F 的节点，$t_{k,s}$ 是方法 S 的节点，则 $t_{k,s}$ 也是方法 F 的节点，在 $t_{k-1,s}$ 和 $t_{k,s}$ 之间有方法 F 的 $r-1$ 个节点．设已用方法 F 和 S 分别积分子系统(11.2)和(11.3)，得到数值解

$$t_{k,F},\ y_{k,F},\ f_{k,F}=f(y_{k,F},\ \tilde{z}_{k,F},\ t_{k,F}),\ k=0,1,\cdots,i,\ i=rj,$$

$$t_{k,s},\ z_{k,s},\ g_{k,s}=g(\tilde{y}_{k,s},z_{k,s},t_{k,s}),\ k=0,\ 1,\cdots j,$$

其中量 $\tilde{y}_{k,s}$，$\tilde{z}_{k,F}$ 按下面的叙述确定．

从节点 $t_{i,F}=t_{j,s}$，假设用方法 F 以步长 h 向前积分(11.2)一步，即先积分快变分量一步．为此，(11.2)中 z 由 $t_{j,s}$ 及其前面的若干个 $t_{k,s}$ 上的已算出的 $z_{k,s}$ 和 $z_{k,s}'=g_{k,s}(k=j,j-1,\cdots)$ 作插值求得．于是得到

$$t_{i+1,F},\ y_{i+1,F},\ f_{i+1,F}=f(y_{i+1,F},\ \tilde{z}(t_{i+1,F}),\ t_{i+1,F}).$$

如此积分 r 步到 $t_{i+r,F}$．然后，用方法 S 向前积分子系统(11.3)一个步长 H．这时，在节点 $t_{k,s}$，$k=j+1,\ j,j-1,\cdots$，上

y的值已经算出，可以使用．用另外的点上的值时，可用插值求得．

下面的算法是计算一个大步长H的完整的循环．设已分别用方法F和S，步长h和H，$H=rh$，积分子系统(11.2)、(11.3)到节点$t_{i,F}$，$t_{j,s}$，且$j=ri$，$t_{i,F}=t_{j,s}$．

算法 I

步 1，$m=1$．

步 2　由$(t_{k,s}, z_{k,s}, g_{k,s})$，$k=i, i-1, \cdots$，构造$z$的插值多项式$\tilde{z}(t)$（例如采用 Nordsieck 的插值表示式），以$\tilde{z}$的值代替z计算函数$f(y, z, t)$所需要的值．

步 3　以步长h，用方法F对子系统(11.2)积分一步．

步 4　$m=m+1$，若$m \leqslant r$，转步 2，若$m>r$，则转步 5．

步 5　由$(t_{k,F}, y_{k,F}, f_{k,F})$，$k=i+r, i+r-1, \cdots$构造$y$的插值多项式$\tilde{y}(t)$，以$\tilde{y}$代$y$计算所需要的函数$g(y, z, t)$的值．

步 6　以步长H，用方法S对子系统(11.3)积分一步．

此时已得到$t_{i+r,F}=t_{j+1,s}$时刻的数值解，即对整个方程(11.1)积分一个大步长H．复重步 1 至步 6 可继续积分．

在 Gear 的程序包[20]中提供了两类解常微分方程组的算法．一类是应用向后微分公式及拟 Newton 迭代方法的算法．它适合于解刚性常微分方程组．另一类是应用通常的 Adams-Moulton 校正公式及简单迭代的算法．它适合于解非刚性方程组．在这个程序包中，两类算法是分别使用的．如果我们把前者看成方法F，而把后者看成方法S，并把它们分别应用于子系统(11.2)和(11.3)就可得到算法I类型的组合方法．这个程序包中贮存每一个节点上的信息的形式为

$$Y_{i,F}=\left[y_{i,F}, \frac{h}{1!}y'_{i,F}, \frac{h^2}{2!}y''_{i,F}, \cdots, \frac{h^k}{k!}y^{(k)}_{i,F}\right],$$

$$Z_{j,s}=\left[z_{j,s}, \frac{H}{1!}z'_{j,s}, \frac{H^2}{2!}z''_{j,s}, \cdots, \frac{H^k}{k!}z^{(k)}_{j,s}\right],$$

其中 k 为所用公式的阶数，插值多项式 $\tilde{y}(t)$ 和 $\tilde{z}(t)$ 可取为 Taylor 展式的形式

$$\tilde{y}(t)=\sum_{p=0}^{k}\left(\frac{t-t_{i,F}}{h}\right)^{p}\frac{h^{p}}{p!}y_{i,F}^{(p)} \tag{11.5}$$

$$\tilde{z}(t)=\sum_{p=0}^{k}\left(\frac{t-t_{j,s}}{H}\right)^{p}\frac{H^{p}}{p!}z_{i,s}^{(p)}, \tag{11.6}$$

当 $H=h$ 时，k 阶积分公式写成求解非线性联立方程组的形式为

$$\begin{aligned}y_{n+1}=&\sum_{i=1}^{k}(\alpha_{F_i}y_{n-i+1}+h\beta_{F_i}y'_{n-i+1})+\\&h\beta_{F_0}f(y_{n+1},z_{n+1},t_{n+1}),\end{aligned} \tag{11.7}$$

$$\begin{aligned}z_{n+1}=&\sum_{i=1}^{k}(\alpha_{s_i}z_{n-i+1}+h\beta_{s_i}z'_{n-i+1})+\\&h\beta_{s_0}g(y_{n+1},z_{n+1},t_{n+1}),\end{aligned} \tag{11.8}$$

其中 α_{F_i}，β_{F_i}，β_{F_0} 为向后微分公式的系数，α_{s_i}，β_{s_i}，β_{s_0} 为 Adams 公式的系数，$i=1,\cdots,k$. 对子系统 (11.7) 采用拟 Newton 迭代（对变量 y），而对子系统 (11.8) 采用简单迭代(对变量 z)。我们得到下面的格式

$$\begin{cases}y_{n+1}^{p+1}=y_{n+1}^{p}+J^{-1}\Big[\sum_{i=1}^{k}(\alpha_{F_i}y_{n-i+1}+h\beta_{F_i}y'_{n-i+1})\\\qquad+h\beta_{F_0}f(y_{n+1}^{p},z_{n+1}^{p},t_{n+1}^{p})-y_{n+1}^{p}\Big],\\z_{n+1}^{p+1}=\sum_{i=1}^{k}(\alpha_{s_i}z_{n-i+1}+h\beta_{s_i}z'_{n-i+1})\\\qquad+h\beta_{s_0}g(y_{n+1}^{p},z_{n+1}^{p},t_{n+1}),\end{cases} \tag{11.9}$$

其中 J 是矩阵 $\left(I-h\beta_{F_0}\dfrac{\partial f}{\partial y}\right)\Big|_{(t_{n+1},\,y_{n+1}^{p},\,z_{n+1}^{p})}$ 的近似，它在积分的若干步内可以保持不变. 迭代 (11.9) 的收敛条件为矩阵

$$\left.\begin{pmatrix}I+J^{-1}\left(h\beta_{F_0}\dfrac{\partial f}{\partial y}-I\right) & h\beta_{F_0}J^{-1}\dfrac{\partial f}{\partial z}\\ h\beta_{s_0}\dfrac{\partial g}{\partial y} & h\beta_{s_0}\dfrac{\partial g}{\partial z}\end{pmatrix}\right|_{(t_{n+1},\,y_{n+1}^{p},\,z_{n+1}^{p})} \tag{11.10}$$

的模小于1．由矩阵 J 的取法，(11.10) 的矩阵近似为

$$\begin{pmatrix} 0 & h\beta_{F_0} J^{-1}\dfrac{\partial f}{\partial z} \\ h\beta_{s_0}\dfrac{\partial g}{\partial y} & h\beta_{s_0}\dfrac{\partial g}{\partial z} \end{pmatrix}$$

因此，若 $\left\|\dfrac{\partial g}{\partial y}\right\|$，$\left\|\dfrac{\partial g}{\partial z}\right\|$ 为通常非刚性方程的量级，而 $\left\|\dfrac{\partial f}{\partial y}\right\| \gg \left\|\dfrac{\partial f}{\partial z}\right\|$ 时，则 h 不需要很小，就可以保证矩阵 (11.10) 的模小于1．从而保证迭代 (11.9) 的收敛性.

根据上面类似的分析，若将向后微分公式同时应用到方程组 (11.2) 和 (11.3)，而在迭代求解非线性联立方程组时将它们再分开，对应于 (11.2) 的方程用拟 Newton 算法，对应于 (11.3) 的方程用简单迭代法，也可收到相当的效果.

由于 $\dfrac{\partial f}{\partial y}$ 的维数一般比起 $\dfrac{\partial H}{\partial x}$ 的维数要小得多，因而用上述的处理可以节省大量的计算量.

若方程组 (11.2)、(11.3) 能写成如下形式

$$y' = Ay + f(z, t), \quad y(0) = y_0, \tag{11.11}$$

$$z' = g(y, z, t), \quad z(0) = z_0, \quad t \in [0, t_e], \tag{11.12}$$

其中 A 为 $m \times m$ 阶常数或分段常数矩阵，$y \in R^m$，$z \in R^n$，即快变子系统可用线性常系数微分方程组来描述．在控制系统的设计中常常可以碰到这种情形．[9] 给出了这种情形的一个组合算法．对于快变子系统 (11.11)，取步长 h 从 t_n 到 t_{n+1}，用隐式 Runge-Kutta 方法积分，例如用二级四阶隐式 Runge-Kutta 公式

$$\begin{array}{c|cc} (3-\sqrt{3})/6 & 1/4 & (3-2\sqrt{3})/12 \\ (3+\sqrt{3})/6 & (3+2\sqrt{3})/12 & 1/4 \\ \hline & 1/2 & 1/2 \end{array} \tag{11.13}$$

应用于 (11.11)，有

$$K_1 = hA\left[y_n + \frac{K_1}{4} + \left(\frac{1}{4} - \frac{\sqrt{3}}{6}\right)K_2\right] + hf,$$

$$K_2 = hA\left[y_n + \left(\frac{1}{4} + \frac{\sqrt{3}}{6}\right)K_1 + \frac{K_2}{4}\right] + hf.$$

这里考虑到(11.12)为慢变子系统，且为了避免叙述的累赘，假定了

$$f(z(t_n + h(3-\sqrt{3})/6), t_n + h(3-\sqrt{3})/6)$$
$$\approx f(z(t_n + h(3+\sqrt{3})/6),$$
$$t_n + h(3+\sqrt{3})/6) \approx f(z(t_n+h), t_n+h) \approx f,$$

解出 K_1, K_2

$$K_1 = \left[I - \frac{hA}{2} + \frac{(hA)^2}{12}\right]^{-1}\left[hA\left(I - \frac{hA\sqrt{3}}{6}\right)\right.$$
$$\left.\cdot y_n + h\left(I - \frac{hA\sqrt{3}}{6}\right)f\right],$$

$$K_2 = \left[I - \frac{hA}{2} + \frac{(hA)^2}{12}\right]^{-1}\left[hA\left(I + \frac{hA\sqrt{3}}{6}\right)\right.$$
$$\left.\cdot y_n + h\left(I + \frac{hA\sqrt{3}}{6}\right)f\right].$$

因此，

$$y_{n+1} = \left[I - \frac{hA}{2} + \frac{(hA)^2}{12}\right]^{-1}$$
$$\cdot \left[\left(I + \frac{hA}{2} + \frac{(hA)^2}{12}\right)y_n + hf\right] \tag{11.14}$$

如此积分 r 个 h 步，即一个步长 H. 然后，用显式 Runge-Kutta 方法积分子系统(11.12)一个大步长 H. 各自所需要的另一个方程组的解分量的值可通过插值得到. 假定 $H = rh$, $H = t_e/N$. 于是，上述的方法可写成如下算法：

算法 II

步 1　$n = 1$;

步 2　$m = 1$;

步 3　由数值解 $(t_{k,s}, y_{k,s}, z_{k,s})$, $k = n, n-1, n-2, \cdots$ 构造 Nordsieck 插值表示式，计算

$$\tilde{z}_{n,m} = z_n + \Delta_m z_n' + \cdots + \Delta_m^{\mu} z_n^{(\mu)}/\mu!,$$

其中 $\Delta_m = t_{n,m} - t_n$, $t_{n,m} = t_n + mh$;

步 4
$$y_{n,m+1} = \left[I - \frac{hA}{2} + \frac{(hA)^2}{12}\right]^{-1} \cdot \left[\left(I + \frac{hA}{2} + \frac{(hA)^2}{12}\right) y_{n,m} + hf(\tilde{z}_{n,m}, t_{n,m})\right];$$

步 5 $m = m + 1$, 若 $m \leqslant r$, 则转步 3,若 $m > r$, 则转步 6;

步 6 由数值解 $(t_{n,F}, y_{n,F}, z_{n,F})$ 和 $y_{n+1,F}$ 构造插值多项式 $\tilde{y}(t)$, 计算 K_1, K_2, K_3, K_4.

其中

$$K_1 = g(\tilde{y}_n, z_n, t_n),$$

$$K_2 = g\left(\tilde{y}_n\left(t_n + \frac{1}{2}H\right), z_n + \frac{H}{2}K_1, t_n + \frac{1}{2}H\right),$$

$$K_3 = g\left(\tilde{y}\left(t_n + \frac{1}{2}H\right), z_n + \frac{H}{2}K_2, t_n + \frac{1}{2}H\right),$$

$$K_4 = g(\tilde{y}(t_n + H), z_n + HK_3, t_n + H);$$

步 7 用步长 H, 用显式 Runge-Kutta 公式

$$z_{n+1} = z_n + \frac{H}{6}(K_1 + 2K_2 + 2K_3 + K_4) \tag{11.15}$$

积分一步;

步 8 $n = n + 1$, 若 $n \leqslant N$, 则转步 2, 若 $n > N$, 则计算终止.

下面我们介绍 Wells[112] 对于系统

$$y' = f(y, z), \quad y(0) = y_0, \tag{11.16a}$$

$$z' = g(y, z), \quad z(0) = z_0, \quad t \in [0, t_e] \tag{11.16b}$$

的组合算法,其中 $y \in R^{d_F}$, $z \in R^{d_s}$. 像前面一样假定 y 是快变系统, z 是慢变系统. 用步长 h 来积分快变分量, 用步长 H 来积分慢变分量, $H = rh$, $H = t_e/N$. 我们分别列出两个算法: 用线性多步方法先积分快变分量的算法 III 和先积分慢变分量的算法 IV:

算法 III 先积分快变分量的算法

步 1 $m = 1$;

步 2 $k = 1$;

步 3 对 $n = (m-1)r + k$ 计算

$$\tilde{z}_n = \sum_{j=1}^{p} [\tilde{\alpha}_{s_k, j} z_{(m-j)r} + H\tilde{\beta}_{s_k, j} z'_{(m-j)r}],$$

$$y_n = \sum_{j=1}^{p} [\alpha_{F_j} y_{n-j} + h\beta_{F_j} y'_{n-j}] + h\beta_{F_0} f(y_n, \tilde{z}_n),$$

$$y'_n = f(y_n, \tilde{z}_n);$$

步 5 $k = k+1$, 若 $k \leqslant r$, 则转步 3, 若 $k > r$, 则转步 6;

步 6
$$z_{mr} = \sum_{j=1}^{p} [\alpha_{s_j} z_{(m-j)r} + H\beta_{s_j} z'_{(m-j)r}] + H\beta_{s_0} g \cdot (y_{mr}, z_{mr}),$$

$$z'_{mr} = g(y_{mr}, z_{mr});$$

步 7 $m = m+1$, 若 $m \leqslant N$, 则转步 2, 若 $m > N$, 则转步 8;

步 8 结束.

这里
$$\alpha_{F_1}, \cdots, \alpha_{F_p}, \ \beta_{F_0}, \cdots, \beta_{F_p}$$
$$\alpha_{s_1}, \cdots, \alpha_{s_p}, \ \beta_{s_0}, \cdots, \beta_{s_p}$$
是线性多步方法的系数. $\tilde{\alpha}_{s_k, j}$ 和 $\tilde{\beta}_{s_k, j}$ 是由慢变分量的预估多项式来确定的. 如果使用 Nordsieck 表示式, $\tilde{z}_n$ 由

$$\tilde{z}_n = z_{(m-1)r} + \rho_n z'_{(m-1)r} + \cdots + \rho_n^{\mu} z^{(\mu)}_{(m-1)r} / \mu!$$

得到,其中 $\rho_n = t_n - t_{(m-1)r}$, μ 是慢变分量预估公式的阶.

类似地利用对快变分量进行外插, 先对慢变分量积分一个大

步长 H，然后利用对慢变分量内插来完成对快变分量积分 r 步，此时有如下算法:

算法 IV. 先积分慢变分量的算法

步 1　$m = 1$;

步 2
$$\bar{y}_{mr} = \sum_{j=0}^{p-1} [\bar{\alpha}_{F_j} y_{(m-1)r-j} + h\bar{\beta}_{F_j} y'_{(m-1)r-j}],$$

$$z_{mr} = \sum_{j=1}^{p} [\alpha_{s_j} z_{(m-j)r} + H\beta_{s_j} z'_{(m-j)r}] + H\beta_{s_0} g(\bar{y}_{mr}, z_{mr}),$$

$$z'_{mr} = g(\bar{y}_{mr}, z_{mr});$$

步 3　$k = 1$;

步 4　对 $n = (m-1)r + k$，计算

$$z_n = \sum_{j=0}^{p-1} [\alpha_{s_{k,j}} z_{(m-j)r} + H\beta_{s_{k,j}} z'_{(m-j)r}],$$

$$y_n = \sum_{j=1}^{p} [\alpha_{F_j} y_{n-j} + h\beta_{F_j} y'_{n-j}] + h\beta_{F_0} f(y_n, z_n),$$

$$y'_n = f(y_n, z_n);$$

步 5　$k = k+1$，若 $k \leqslant r$，则转步 4，若 $k > r$ 则转步 6;

步 6　$m = m+1$，若 $m \leqslant N$，则转步 2，若 $m > N$，则转步 7;

步 7　结束.

这里的 $\bar{\alpha}_{F_j}$ 和 $\bar{\beta}_{F_j}$ 是由快变系统预估多项式确定的，$\alpha_{s_{k,j}}$ 和 $\beta_{s_{k,j}}$ 是由慢变系统校正多项式确定的.

如果用 Nordsieck 表示式，快变分量的外插公式为

$$\bar{y}_{mr} = y_{(m-1)r} + Hy'_{(m-1)r} + \cdots + H^\nu y^{(\nu)}_{(m-1)r}/\nu!,$$

慢变分量的内插公式为

$$z_n = z_{mr} + \rho_n z'_{mr} + \cdots + \rho_n^\mu z^{(\mu)}_{mr}/\mu!,$$

$$\rho_n = t_n - t_{mr}.$$

算法 III 和 IV 中的隐式方程，如果子系统是非刚性方程组，可以用简单迭代法求解，如果子系统是刚性方程组，可以用拟 Ne-

wton 迭代法求解．按 Gear 的想法，不同变化速率系统最好按解分量变化的快慢分成多个组，用不同的方法处理．他认为，最终应当按解的变化自动地分组，不要人去干预．但是，这样花费在分组处理上的计算太大，尚难得到好用的方法．

由于在快变和慢变子系统中，可以使用不同的数值方法，如果两个子系统之间耦合比较小，使用组合算法计算量可以减少很多．一般 Newton 迭代每一步需要求解线性代数方程组 $wx = b$．用 LU 分解求解这个方程组，需要 $k\cdot(\dim(x))^2$ 个运算，其中 $\dim(x)$ 是 x 的维数，k 是比例常数．若快变子系统和慢变子系统的维数分别为 d_F 和 d_s，使用组合算法，用 $H = rh$ 积分一步只需要 $k(rd_F^2 + d_s^2)$ 个运算．整个系统 (11.1) 用刚性方法的算法，用步长 h 积分 r 步则需要 $kr(d_F + d_s)^2$个运算．因此，组合算法节省 $k[(r-1)d_s^2 + 2rd_Fds]$ 个运算．d_s 越大节省越多，即使 $r=2$，节省的计算量也是相当可观的．

§3 方法的收敛性和误差阶

Gear[57] 和 Wells[114] 讨论了用线性多步方法构造的组合方法的收敛性和误差阶．这里我们摘录 [3] 中讨论的用单步法构造的组合方法的收敛性和误差阶． 子系统 (11.2) 和 (11.3) 可以写为

$$\frac{dy(t)}{dt} = f(y(t), z(t), t),\quad y(t_0) = y_0, \tag{11.17}$$

$$\frac{dz(t)}{dt} = g(y(t), z(t), t),\quad z(t_0) = z_0, \tag{11.18}$$

对 (11.17)、(11.18) 作下面的假定：

假定 11.1 函数 f, g 在 $[t_0, T]$上对 t 连续，而对 y 和 z 满足 Lipschitz 条件，即存在正常数 L_{f1}, L_{f2}, L_{g1}, L_{g2}，对任意的 $y^{(1)}$, $y^{(2)}\in E_m$ 和 $z^{(1)}$, $z^{(2)}\in E_n$ 有

$$\|f(y^{(1)}, z^{(1)}, t) - f(y^{(2)}, z^{(1)}, t)\| \leqslant L_{f1}\|y^{(1)} - y^{(2)}\|,$$

$$\|f(y^{(1)}, z^{(1)}, t) - f(y^{(1)}, z^{(2)}, t)\| \leqslant L_{f2}\|z^{(1)} - z^{(2)}\|,$$

$$\|g(y^{(1)}, z^{(1)}, t) - g(y^{(2)}, z^{(1)}, t)\| \leqslant L_{g1}\|y^{(1)} - y^{(2)}\|,$$
$$\|g(y^{(1)}, z^{(1)}, t) - g(y^{(1)}, z^{(2)}, t)\| \leqslant L_{g2}\|z^{(1)} - z^{(2)}\|.$$

在这个假定下，(11.17)、(11.18) 在 $[t_0, T]$ 上有唯一的连续可微解，记其解为 $\bar{y}(t)$, $\bar{z}(t)$.

为了讨论方便，这里我们还需要描述用一般的单步方法数值求解 (11.17)、(11.18) 所构造的组合算法的模型. 用点 $t_i = t_0 + ih_y$, $i = 0, 1, \cdots, N$ 将 $[t_0, T]$ 离散化. 用点 $t_{ij} = t_{i-1} + jh_z$, $j = 0, 1, \cdots, M$ 将区间 $[t_{i-1}, t_i]$ 离散化, $t_{i0} = t_{i-1}$, $t_{iM} = t_i$. 点 t_i 和 t_{ij} 称作离散化节点, h_y, h_z 为步长. 为简单起见，只考虑定步长. 设已得到序列 y_k, dy_k, $k = 0, 1, \cdots, i$ 和 z_{lj}, dz_{lj}, $l = 1, \cdots, i$, $j = 0, 1, \cdots, M$, 并记 $z_l = z_{lM}$, 其中 y_k, dy_k 分别是 $\bar{y}(t_k)$, $\dot{\bar{y}}(t_k)$ 的近似值, z_{lj}, dz_{lj} 分别是 $\bar{z}(t_{lj})$, $\dot{\bar{z}}(t_{lj})$ 的近似值.这里我们取 $dy_k = f(y_k, z_k, t_k)$, $dz_{lj} = g(\tilde{y}_l(t_{lj}), z_{lj}, t_{lj})$, 其中 $\tilde{y}_l(t)$ 在下面定义.

记

$$T_{y_i} = (t_0, t_1, \cdots, t_i)^T, \quad T_{z_i} = (t_{10}, t_{11}, \cdots, t_{iM})^T.$$
$$Y_i^T = (y_0^T, y_1^T, \cdots, y_i^T), \quad DY_i^T = (dy_0^T, dy_1^T, \cdots, dy_i^T),$$
$$Z_i^T = (z_{10}^T, z_{11}^T, \cdots, z_{iM}^T), \quad DZ_i^T = (dz_{10}^T, dz_{11}^T, \cdots, dz_{iM}^T),$$

并记它们所属的向量空间为 $E^{T_{y_i}}, E^{T_{z_i}}, E_{Y_i}, E_{z_i}$, 即有 $T_{y_i} \in E^{T_{y_i}}$, $T_{z_i} \in E^{T_{z_i}}$, Y_i, $DY_i \in E_{Y_i}$, Z_i, $DZ_i \in E_{Z_i}$. 构造分别由 (Y_i, DY_i, T_{y_i}) 和 (Z_i, DZ_i, T_{z_i}) 确定的插值函数 $I_{y_i}(Y_i, DY_i, T_{y_i})(t)$ 和 $I_{z_i}(Z_i, DZ_i, T_{z_i})(t)$, 它们分别为由$[t_0, T]$到 E_m 和 E_n 中的函数. 令

$$\tilde{y}_i(t) = I_{y_i}(Y_i, DY_i, T_{y_i})(t), \tag{11.19}$$
$$\tilde{z}(t) = I_{z_i}(Z_i, DZ_i, T_{Z_i})(t). \tag{11.20}$$

假定在构造 $\tilde{y}_i(t)$ 时，插值函数 $I_{y_i}(Y_i, DY_i, T_{y_i})(t)$ 未用 dy_i, 并且对于 $l \leqslant i$ 有 $\tilde{y}_i(t_l) = y_l$, $\tilde{z}(t_{lj}) = z_{lj}$. 将 (11.20) 代入 (11.17) 中的 $z(t)$, 得到初值问题

$$\frac{dy(t)}{dt} = f(y(t), \tilde{z}_i(t), t), \quad y(t_i) = y_i, \tag{11.21}$$

用单步公式以步长 h_y 数值积分一步,得

$$y_{i+1} = y_i + h_y\phi_f(y_i, \tilde{z}_i, t_i, h_y), \tag{11.22}$$

其中增量函数 ϕ_f 是由所用的单步公式和 f 确定的,仅是 y_i, $\tilde{z}_i(t)$, t_i, h_y 的函数. 由 Y_{i+1}, DY_{i+1}, $T_{y_{i+1}}$ 按(11.19)式构造 $\tilde{y}_{i+1}(t) = I_{y_{i+1}}(Y_{i+1}, DY_{i+1}, T_{y_{i+1}})(t)$, 并代入(11.18)中的 $y(t)$, 得到初值问题

$$\frac{dz(t)}{dt} = g(\tilde{y}_{i+1}(t), z(t), t), \quad z(t_{i+1,0}) = z_{iM} = z_i. \tag{11.23}$$

用单步公式以步长 h_z 数值积分,得到递推公式

$$z_{i+1,j+1} = z_{i+1,j} + h_z\phi_g(z_{i+1,j}, \tilde{y}_{i+1}, t_{i+1,j}, h_z),$$
$$j = 0, 1, \cdots, M-1, \quad z_{i+1,0} = z_{iM_0}, \tag{11.23$_1$}$$

其中增量函数 ϕ_g 是由所用的单步公式和 g 确定的,仅是 $z_{i+1,j}$, $\tilde{y}_{i+1}(t)$, $t_{i+1,j}$, h_z 的函数.

这样,我们描述了一个从 t_i 推进到 t_{i+1} 的完整过程. 将公式(11.22)和(11.23$_1$)中的 i 用 $i+1$ 代替,不断重复这个过程,直到 $t_i = T$ 为止.

为了证明算法的收敛性,还需要一些假定和处理. 设量 e_i 满足式

$$\|e_{i+1}\| \leqslant (1 + hL)\|e_i\| + D, \quad i = 0, 1, \cdots. \tag{11.24}$$

文[20]给出下面的引理.

引理 11.1 如果 $\|e_i\|$ 满足(11.24), 并且 $0 \leqslant ih \leqslant b$, 则

$$\|e_i\| \leqslant \frac{e^{Lb} - 1}{hL} D + e^{Lb}\|e_0\|. \tag{11.25}$$

对(11.17)、(11.18)的解 $\bar{y}(t)$, $\bar{z}(t)$ 令

$$\begin{cases} \bar{Y}_i^T = (\bar{y}(t_0)^T, \bar{y}(t_1)^T, \cdots, \bar{y}(t_i)^T), \ D\bar{Y}_i^T = (\dot{\bar{y}}(t_0)^T, \\ \qquad \dot{\bar{y}}(t_1)^T, \cdots, \dot{\bar{y}}(t_i)^T), \\ \bar{Z}_i^T = (\bar{z}(t_{10})^T, \bar{z}(t_{11})^T, \cdots, \bar{z}(t_{iM})^T), \ D\bar{Z}_i^T = (\dot{\bar{z}}(t_{10})^T, \\ \qquad \dot{\bar{z}}(t_{11})^T, \cdots, \dot{\bar{z}}(t_{iM})^T). \end{cases} \tag{11.26}$$

记

$$\varepsilon_{inty} = \max_i\left(\max_{t_0 \leqslant t \leqslant t_i} \|\bar{y}(t) - I_{y_i}(\bar{Y}_i, \overline{DY}_i, T_{y_i})(t)\|\right), \tag{11.27}$$

$$e_{\text{intz}} = \max_i(\max_{t_0 \leqslant t \leqslant t_i + h_y} \|\bar{z}(t) - I_{z_i}(\bar{Z}_i, D\bar{Z}_i, T_{z_i})(t)\|).$$

(11.28)

假定 11.2 (11.19)、(11.20) 的映象 $I_{y_i}(\cdot, \cdot, \cdot)(t)$ 和 $I_{z_i}(\cdot,\cdot,\cdot)(t)$ 满足下面的条件:

(i) $\lim\limits_{h_y \to 0} e_{\text{inty}} = 0$, $\lim\limits_{\substack{h_y \to 0 \\ h_z \to 0}} e_{\text{intz}} = 0$;

(ii) 存在正常数 L_{y1}, L_{y2} 和 L_{z1}, L_{z2}, 使对任何 $(Y_i^{(1)}, DY_i^{(1)}) \in E_{Y_i} \times E_{Y_i}$, $(Y_i^{(2)}, DY_i^{(2)}) \in E_{Y_i} \times E_{Y_i}$ 和 $(Z_i^{(1)}, DZ_i^{(1)}) \in E_{E_i} \times E_{Z_i}$, $(Z_i^{(2)}, DZ_i^{(2)}) \in E_{Z_i} \times E_{Z_i}$ 成立估计式

$$\|I_{y_i}(Y_i^{(1)}, DY_i^{(1)}, T_{y_i})(t) - I_{y_i}(Y_i^{(2)}, DY_i^{(2)}, T_{y_i})(t)\|$$
$$\leqslant L_{y1}\max_{j \leqslant i}\|y_j^{(1)} - y_j^{(2)}\| + L_{y2}\max_{j \leqslant i-1}\|\mathrm{d}y_j^{(1)} - dy_j^{(2)}\|,$$
$$t \in [t_0, t_i],$$

$$\|I_{z_i}(Z_i^{(1)}, DZ_i^{(1)}, T_{z_i})(t) - I_{z_i}(Z_i^{(2)}, DZ_i^{(2)}, T_{z_i})(t)\|$$
$$\leqslant L_{z1}\max_{l \leqslant i,\, j \leqslant M}\|z_{lj}^{(1)} - z_{lj}^{(2)}\| + L_{z2}\max_{l \leqslant i,\, j \leqslant M}\|dz_{lj}^{(1)} - dz_{lj}^{(2)}\|,$$
$$t \in [t_0, t_i + h_y].$$

(iii) 若 Y_{i-1}, DY_{i-1} 分别是由 Y_i 和 DY_i 去掉 y_i 和 dy_i 得到的向量,则当 $t \leqslant t_{i-1}$ 时,有

$$I_{y_i}(Y_i, DY_i, T_{y_i})(t) = I_{y_{i-1}}(Y_{i-1}, DY_{i-1}, T_{y_{i-1}})(t).$$

由这个假定的 (iii) 可以看出, 分别将 $\tilde{y}_{i-1}(t) = I_{y_{i-1}}(Y_{i-1}, DY_{i-1}, T_{y_{i-1}})(t)$ 和 $\tilde{y}_i(t) = I_{y_i}(Y_i, DY_i, T_{y_i})(t)$ 代入(11.18), 得到的两个微分方程组的初值问题的解当 $t \leqslant t_{i-1}$ 是重合的. 对常微分方程初值问题

$$\frac{dz(t)}{dt} = g(\tilde{y}_{i+1}(t), z(t), t), \quad z(t_0) = z_0. \tag{11.29}$$

用步长 h_z 的递推公式

$$z_{l,j+1} = z_{lj} + h_z\phi_g(z_{lj}, \tilde{y}_{i+1}, t_{lj}, h_z), \; z_{10} = z_0,$$
$$z_{l+1,0} = z_{lM} \tag{11.30}$$

数值积分得到的 z_{lj}, $l \leqslant i+1$, $j = 0, 1, \cdots, M$, 将与前面所述的算法得到的 E_{lj} 是重合的.

记 $\Delta_f(\bar{y}(t), \bar{z}, t, h)$ 是 (11.17) 的精确增量函数，即有

$$\bar{y}(t+h) = \bar{y}(t) + h\Delta_f(\bar{y}(t), \bar{z}, t, h),$$

其中 $\bar{y}(t)$ 和 $\bar{z} = \bar{z}(t)$ 是 (11.17)、(11.18) 的精确解. 记

$$e_{\psi f}(t, h) = \Delta_f(\bar{y}(t), \bar{z}, t, h) - \psi_f(\bar{y}(t), \bar{z}, t, h),$$

$$\bar{e}_{\psi f}(h) = \max_{t_0 \leqslant t \leqslant T} \|e_{\psi f}(t, h)\|.$$

对增量函数 $\psi_f(y, z, t, h)$，作下面的假定.

假定 11.3 增量函数 $\psi_f(y, z, t, h)$ 满足下面的条件:

(i) 在所考虑的区域中，$\psi_f(y, z, t, h)$ 对于 t, h 连续，对 y, z 满足 Lipschitz 条件，即存在正常数 $L_{\psi f1}$, $L_{\psi f2}$ 和 h_{y_0}，使对任意的 $y^{(1)}$, $y^{(2)}$ 和连续函数 $z^{(1)}(s)$, $z^{(2)}(s)$，当 $h \leqslant h_{y_0}$ 时，成立

$$\|\psi_f(y^{(1)}, z^{(1)}, t, h) - \psi_f(y^{(2)}, z^{(1)}, t, h)\| \leqslant L_{\psi f1}\|y^{(1)} - y^{(2)}\|,$$

$$\|\psi_f(y^{(1)}, z^{(1)}, t, h) - \psi_f(y^{(1)}, z^{(2)}, t, h)\| \leqslant L_{\psi f2} \max_{t \leqslant s \leqslant t+h} \|z^{(1)}(s) - z^{(2)}(s)\|.$$

(ii) $\lim_{h \to 0} \bar{e}_{\psi f}(h) = 0$.

类似地，对函数 $\bar{y}(t)$, $\bar{z}(t)$ 和增量函数 $\psi_g(z, y, t, h)$ 定义 $\Delta_g(\bar{z}(t), \bar{y}, t, h)$ 和 $e_{\psi g}(t, h)$, $\bar{e}_{\psi g}(h)$，并作如下假定:

假定 11.4 增量函数 $\psi_g(z, y, t, h)$ 满足下面的条件:

(i) 在所考虑的区域中，$\psi_g(z, y, t, h)$ 对 t, h 连续，对 z, y 满足 Lipschitz 条件，即存在正常数 $L_{\psi g1}$, $L_{\psi g2}$ 和 h_{z_0} 使对任意的 $z^{(1)}$, $z^{(2)}$ 和连续函数 $y^{(1)}(s)$, $y^{(2)}(s)$，当 $h \leqslant h_{z_0}$ 时，成立

$$\|\psi_g(z^{(1)}, y^{(1)}, t, h) - \psi_g(z^{(2)}, y^{(1)}, t, h)\| \leqslant L_{\psi g1}\|z^{(1)} - z^{(2)}\|,$$

$$\|\psi_g(z^{(1)}, y^{(1)}, t, h) - \psi_g(z^{(1)}, y^{(2)}, t, h)\| \leqslant L_{\psi g2} \max_{t \leqslant s \leqslant t+h} \|y^{(1)}(s) - y^{(2)}(s)\|.$$

(ii) $\lim_{h \to 0} \bar{e}_{\psi g}(h) = 0$.

下面两个定理分别给出算法的收敛性和收敛阶.

定理 11.1 设假定 11.1、11.2、11.3、11.4 成立，则当 $h_y \to 0$

时，由组合算法构造的序列 $\{y_i\}$，$\{z_i\}$ 将收敛到 (11.17)、(11.18) 两方程组的解 $\bar{y}(t)$，$\bar{z}(t)$，即当 $h_y \to 0$ 时，若有 $t_i \to t$，则 $y_i \to \bar{y}(t)$，$z_i \to \bar{z}(t)$。

证明　记 $u_{y_i} = \bar{y}(t_i) - y_i$，$u_{z_{ij}} = \bar{z}(t_{ij}) - z_{ij}$，$u_{z_i} = \bar{z}(t_i) - z_i$。由 $\tilde{z}_i(s)$ 的定义 (11.20) 和假定 11.2，有

$$\begin{aligned}\|\bar{z}(s) - \tilde{z}_i(s)\| &\leqslant \|\bar{z}(s) - I_{z_i}(\bar{Z}_i, D\bar{Z}_i, T_{z_i})(s)\| \\ &\quad + \|I_{z_i}(\bar{Z}_i, D\bar{Z}_i, T_{z_i})(s) - I_{z_i}(Z_i, DZ_i, T_{z_i})(s)\| \\ &\leqslant e_{\text{intz}} + L_{z1}\max_{l\leqslant i, j\leqslant M} \|\bar{z}(t_{lj}) - z_{lj}\| \\ &\quad + L_{z2}\max_{l\leqslant i, j\leqslant M} \|\dot{z}(t_{lj}) - dz_{lj}\|. \end{aligned} \tag{11.31}$$

同样，对 $y_{i+1}(s)$ 有

$$\begin{aligned}\|\bar{y}(s) - \tilde{y}_{i+1}(s)\| &\leqslant \|\bar{y}(s) - I_{y_{i+1}}(\bar{Y}_{i+1}, D\bar{Y}_{i+1}, T_{y_{i+1}})(s)\| \\ &\quad + \|I_{y_{i+1}}(\bar{Y}_{i+1}, D\bar{Y}_{i+1}, T_{y_{i+1}})(s) - I_{y_{i+1}}(Y_{i+1}, DY_{i+1}, T_{y_{i+1}})(s)\| \\ &\leqslant e_{\text{inty}} + L_{y1}\max_{j\leqslant i+1} \|\bar{y}(t_j) - y_j\| + L_{y2}\max_{j\leqslant i} \|\dot{y}(t_j) - dy_j\|. \end{aligned} \tag{11.32}$$

根据前面对 dy_j 的取法，

$$\begin{aligned}\|\dot{y}(t_j) - dy_j\| &= \|f(\bar{y}(t_j), z(t_j), t_j) - f(y_j, \tilde{z}_j(t_j), t_j)\| \\ &\leqslant L_{f1}\|\bar{y}(t_j) - y_j\| + L_{f2}\|\bar{z}(t_j) - z_j\|. \end{aligned} \tag{11.33}$$

将其代入 (11.32)，得

$$\begin{aligned}\|\bar{y}(s) - \tilde{y}_{i+1}(s)\| &\leqslant e_{\text{inty}} + (L_{y1} + L_{y2}L_{f1})\max_{j\leqslant i+1} \|u_{y_j}\| \\ &\quad + L_{y2}L_{f2}\max_{j\leqslant i}\|u_{z_j}\|. \end{aligned} \tag{11.34}$$

由于当 $l \leqslant i$ 时

$$\begin{aligned}\|\dot{z}(t_{lj}) - dz_{lj}\| &= \|g(\bar{y}(t_{lj}), \bar{z}(t_{lj}), t_{lj}) - g(\tilde{y}_i(t_{lj}), z_{lj}), z_{lj})\| \\ &\leqslant L_{g1}\|\bar{y}(t_{lj}) - \tilde{y}(t_{lj})\| + L_{g2}\|\bar{z}(t_{lj}) - z_{lj}\|, \end{aligned} \tag{11.35}$$

将 (11.34) 中的 $i+1$ 改成 i，并代入上式，当 $l \leqslant i$ 时，得

$$\begin{aligned}\|\dot{z}(t_{lj}) - dz_{lj}\| &\leqslant L_{g1}e_{\text{inty}} + L_{g1}(L_{y1} + L_{y2}L_{f1})\max_{j\leqslant i}\|u_{y_j}\| \\ &\quad + L_{g1}L_{y2}L_{f2}\max_{j\leqslant i-1} \|u_{z_j}\| + L_{g2}\|u_{z_{lj}}\|. \end{aligned} \tag{11.36}$$

将 (11.36) 代入 (11.31)，得

$$\|\bar{z}(s) - \tilde{z}_i(s)\| \leqslant e_{\text{intz}} + L_{z1}\max_{l\leqslant i, j\leqslant M} \|u_{z_{lj}}\| + L_{z2}L_{g1}e_{\text{inty}}$$

$$+L_{z2}L_{g1}(L_{y1}+L_{y2}L_{f1})\max_{j\leqslant i}\|u_{y_j}\|+L_{z2}L_{g1}L_{y2}L_{f2}\max_{j\leqslant i-1}\|u_{z_j}\|$$

$$+L_{z2}L_{g2}\max_{l\leqslant i,\ j\leqslant M}\|u_{z_{lj}}\|\leqslant e_{\text{int}z}+L_{z2}L_{g1}e_{\text{int}y}$$

$$+L_{z2}L_{g1}(L_{y1}+L_{y2}L_{f1})\max_{j\leqslant i}\|u_{y_j}\|+(L_{z1}+L_{z2}L_{g2}$$

$$+L_{z2}L_{g1}L_{y2}L_{f2})\max_{l\leqslant i,\ j\leqslant M}\|u_{z_{lj}}\|. \tag{11.37}$$

由假定 11.2 后的说明，当 $l\leqslant i+1$ 时，z_{lj} 满足递推公式

$$z_{l,j+1}=z_{lj}+h_z\phi_g(z_{lj},\tilde{y}_{i+1},t_{lj},h_z),\ z_{l0}=z_{l-1,M},\ l\leqslant i+1,\ j\leqslant M-1.$$

于是

$$\begin{aligned}u_{z_{l,j+1}}&=\bar{z}(t_{l,j+1})-z_{l,j+1}\\&=\bar{z}(t_{lj})+h_z\Delta_g(\bar{z}(t_{lj}),\bar{y},t_{lj},h_z)-z_{lj}\\&\quad-h_z\phi_g(z_{lj},\tilde{y}_{i+1},t_{lj},h_z)\\&=u_{z_{lj}}+h_z[\Delta_g(\bar{z}(t_{lj}),\bar{y},t_{lj},h_z)-\phi_g(\bar{z}(t_{lj}),\bar{y},t_{lj},h_z)]\\&\quad+h_z[\phi_g(\bar{z}(t_{lj}),\bar{y},t_{lj},h_z)-\phi_g(z_{lj},\bar{y},t_{lj},h_z)]\\&\quad+h_z[\phi_g(z_{lj},\bar{y},t_{lj},h_z)-\phi_g(z_{lj},\tilde{y}_{i+1},t_{lj},h_z)].\end{aligned} \tag{11.38}$$

由定理的假定，得到估计式

$$\|u_{z_{l,j+1}}\|\leqslant\|u_{z_{lj}}\|+h_z\bar{e}_{\phi g}(h_z)+h_zL_{\phi g1}\|u_{z_{lj}}\|+h_zL_{\phi g2}\max_{t_0\leqslant s\leqslant t_{i+1}}\|\bar{y}(s)-\tilde{y}_{i+1}(s)\|. \tag{11.39}$$

将 (11.34) 代入，得

$$\begin{aligned}\|u_{z_{l,j+1}}\|&\leqslant(1+h_zL_{\phi g1})\|u_{z_{lj}}\|+h_z\bar{e}_{\phi g}(h_z)+h_zL_{\phi g2}e_{\text{int}y}\\&\quad+h_zL_{\phi g2}(L_{y1}+L_{y2}L_{f1})\max_{j\leqslant i+1}\|u_{y_j}\|\\&\quad+h_zL_{\phi g2}L_{y2}L_{f2}\max_{j\leqslant i}\|u_{z_j}\|.\end{aligned} \tag{11.40}$$

令 $l=i+1$，对 $j=0,1,\cdots,M$，由引理 11.1，我们得到

$$\begin{aligned}\|u_{z_{i+1,j}}\|&\leqslant\frac{e^{L_{\phi g1}h_y}-1}{L_{\phi g1}}[\bar{e}_{\phi g}(h_z)+L_{\phi g2}e_{\text{int}y}+L_{\phi g2}(L_{y1}\\&\quad+L_{y2}L_{f1})\max_{j\leqslant i+1}\|u_{y_j}\|+L_{\phi g2}L_{y2}L_{f2}\max_{j\leqslant i}\|u_{z_j}\|]\\&\quad+e^{L_{\phi g1}h_y}\|u_{z_i}\|\end{aligned}$$

$$= \frac{e^{L_{\psi g1}h_y}-1}{L_{\psi g1}} L_{\psi g2}L_{y2}L_{f2}\max_{j<i}\|u_{z_j}\| + e^{L_{\psi g1}h_y}\|u_{z_i}\|$$

$$+ \frac{e^{L_{\psi g_1}h_y}-1}{L_{\psi g1}} L_{\psi g2}(L_{y1}+L_{y2}L_{f1})\max_{j\leqslant i+1}\|u_{y_j}\|$$

$$+ \frac{e^{L_{\psi g1}h_y}-1}{L_{\psi g1}} [\bar{e}_{\psi g}(h_z) + L_{\psi g2}e_{\text{inty}}]. \qquad (11.41)$$

特别，当 $i=M$ 时，此式也成立，得

$$\|u_{z_{i+1}}\| \leqslant \frac{e^{L_{\psi g1}h_y}-1}{L_{\psi g1}} L_{\psi g2}L_{y2}L_{f2}\max_{j<i}\|u_{z_j}\| + e^{L_{\psi g1}h_y}\|u_{z_i}\|$$

$$+ \frac{e^{L_{\psi g1}h_y}-1}{L_{\psi g1}} L_{\psi g2}(L_{y1}+L_{y2}L_{f1})\max_{j\leqslant i+1}\|u_{y_j}\|$$

$$+ \frac{e^{L_{\psi g1}h_y}-1}{L_{\psi g1}} [\bar{e}_{\psi g}(h_z) + L_{\psi g2}e_{\text{inty}}]. \qquad (11.42)$$

由于

$$\begin{aligned} u_{y_{i+1}} &= \bar{y}(t_{i+1}) - y_{i+1} \\ &= \bar{y}(t_i)+h_y\Delta_f(\bar{y}(t_i), \bar{z}, t_i, h_y)-y_i-h_y\phi_f(y_i, \tilde{z}_i, t_i, h_y) \\ &= u_{y_i} + h_y[\Delta_f(\bar{y}(t_i), \bar{z}, t_i, h_y) - \phi_f(\bar{y}(t_i), \bar{z}, t_i, h_y)] \\ &\quad + h_y[\phi_f(\bar{y}(t_i), \bar{z}, t_i, h_y) - \phi_f(y_i, \bar{z}, t_i, h_y)] \\ &\quad + h_y[\phi_f(y_i, \bar{z}, t_i, h_y) - \phi_f(y_i, \tilde{z}_i, t_i, h_y)], \end{aligned} \qquad (11.43)$$

应用定理的假定，得

$$\begin{aligned} \|u_{y_{i+1}}\| \leqslant{}& \|u_{y_i}\| + h_y\bar{e}_{\psi f}(h_y) + h_yL_{\psi f1}\|u_{y_i}\| \\ &+ h_yL_{\psi f2}\max_{t_i\leqslant s\leqslant t_{i+1}}\|z(s)-\tilde{z}_i(s)\|. \end{aligned} \qquad (11.44)$$

将 (11.37) 代入，得

$$\begin{aligned} \|u_{y_{i+1}}\| \leqslant{}& (1+h_yL_{\psi f1})\|u_{y_i}\| + h_y\bar{e}_{\psi f}(h_y) + h_yL_{\psi f2}e_{\text{intz}} \\ &+h_yL_{\psi f2}L_{z2}L_{g1}e_{\text{inty}}+h_yL_{\psi f2}L_{z2}L_{g1}(L_{y1}+L_{y2}L_{f1})\max_{j<i}\|u_{y_j}\| \\ &+h_yL_{\psi f2}(L_{z1}+L_{z2}L_{g2}+L_{z2}L_{g1}L_{y2}L_{f2})\max_{l\leqslant i,j\leqslant M}\|u_{z_{lj}}\| \\ ={}& (1+h_yL_{\psi f1})\|u_{y_i}\|+h_yL_{\psi f2}L_{z2}L_{g1}(L_{y1}+L_{y2}L_{f1})\max_{j<i}\|u_{y_j}\| \\ &+ h_yL_{\psi f2}(L_{z1}+L_{z2}L_{g2}+L_{z2}L_{g1}L_{y2}L_{f2})\max_{l\leqslant i,j\leqslant M}\|u_{z_{lj}}\| \end{aligned}$$

$$+ h_y[\bar{e}_{\psi f}(h_y) + L_{\psi f2}e_{\text{int}z} + L_{\psi f2}L_{z2}L_{g1}e_{\text{int}y}]. \tag{11.45}$$

令 w_{y_i}, w_{z_i} 满足递推式

$$\begin{aligned} w_{y_{i+1}} &= A_{11}w_{y_i} + A_{12}w_{z_i} + B_1, \ w_{y_0} = 0, \\ w_{z_{i+1}} &= A_{21}w_{y_i} + A_{22}w_{z_i} + B_2, \ w_{z_0} = 0, \end{aligned} \tag{11.46}$$

其中

$$\begin{cases} A_{11} = 1 + h_y[L_{\psi f1} + L_{\psi f2}L_{z2}L_{g1}(L_{y1} + L_{y2}L_{f1})], \\ A_{12} = h_y L_{\psi f2}(L_{z1} + L_{z2}L_{g2} + L_{z2}L_{g1}L_{y2}L_{f2}), \\ A_{21} = \dfrac{e^{L_{\psi g1}h_y} - 1}{L_{\psi g1}} L_{\psi g2}(L_{y1} + L_{y2}L_{f1})A_{11}, \\ A_{22} = e^{L_{\psi g1}h_y} + \dfrac{e^{L_{\psi g1}h_y} - 1}{L_{\psi g1}}[L_{\psi g2}L_{y2}L_{f2} + L_{\psi g2}(L_{y1} \\ \qquad + L_{y2}L_{f1})A_{12}], \\ B_1 = h_y[\bar{e}_{\psi f}(h_y) + L_{\psi f2}e_{\text{int}z} + L_{\psi f2}L_{z2}L_{g1}e_{\text{int}y}], \\ B_2 = \dfrac{e^{L_{\psi g1}h_y} - 1}{L_{\psi g1}}[\bar{e}_{\psi g}(h_z) + L_{\psi g2}e_{\text{int}y} + L_{\psi g2}(L_{y1} \\ \qquad + L_{y2}L_{f1})B_1]. \end{cases} \tag{11.47}$$

由 (11.42)、(11.45), 并将 (11.45) 代入 (11.42) 的中 $\|u_{y_{i+1}}\|$, 容易看出有

$$\begin{aligned} &\|u_{y_i}\| \leqslant w_{y_i}, \\ &\|u_{z_{ij}}\| \leqslant w_{z_i}, \ i = 0, 1, \cdots, N, \ j = 0, 1, \cdots, M, \end{aligned} \tag{11.48}$$

由 A_{11}, A_{12}, A_{21}, A_{22} 的表达式可知它们可表示成

$$A_{11} = 1 + h_y a_{11}$$

$$A_{12} = h_y a_{12}$$

$$A_{21} = h_y a_{21}$$

$$A_{22} = 1 + h_y a_{22}$$

其中 a_{11}、a_{12}、a_{21}、a_{22} 均是 h_y 的正函数. 当 $h_y \to 0$ 时, 均以正数为极限. 可找到正数 $\bar{h}_y$ 和正常数 $\bar{a}_{11}$、$\bar{a}_{12}$、$\bar{a}_{21}$、$\bar{a}_{22}$, 使得当 $h_y \leqslant \bar{h}_y$ 时有 $a_{11} \leqslant \bar{a}_{11}$, $a_{12} \leqslant \bar{a}_{12}$, $a_{21} \leqslant \bar{a}_{21}$, $a_{22} \leqslant \bar{a}_{22}$. 设量 w_i 满足递推式

$$w_{i+1} = (1 + h_y\bar{a})w_i + B, \ w_0 = 0 \tag{11.49}$$

其中 $$\bar{a}=\max\{\bar{a}_{11}+\bar{a}_{12},\ \bar{a}_{21}+\bar{a}_{22}\}$$

$$B=\max\{B_1,\ B_2\}$$

则易证有 $\max\{w_{y_i},\ w_{z_i}\}\leqslant w_i$. 应用引理 11.1, w_i 有估计式

$$w_i\leqslant\frac{e^{\bar{a}(t_i-t_0)}-1}{\bar{a}}\frac{B}{h_y}\tag{11.50}$$

由 B_1, B_2 的表示式及定理的假定易知当 $h_y\to 0$ 时, $\frac{1}{h_y}B\to 0$.

由 (11.50) 知当 t_i 小于某个有限数 $\bar{t}$ 时,一致地有 $\|\bar{y}(t_i)-y_i\|\to 0$, $\|\bar{z}(t_i)-z_i\|\to 0$. 于是, 若有 $h_y\to 0$ 和 $t_i\to t$, 由于 $\|\bar{y}(t)-y_i\|\leqslant\|\bar{y}(t)-\bar{y}(t_i)\|+\|\bar{y}(t_i)-y_i\|$ 和 $y(t)$ 的连续性,将有 $\|\bar{y}(t)-y_i\|\to 0$, 即有 $y_i\to\bar{y}(t)$. 同理可证 $z_i\to\bar{z}(t)$. 定理证毕.

定理 11.1 给出了组合算法的收敛条件.在对$\bar{e}_{\psi f}(h_y)$,$\bar{e}_{\psi g}(h_z)$, e_{inty}, e_{intz} 的更精细的假定下,下面的定理给出了当 $h_y\to 0$ 时, $y_i\to\bar{y}(t)$, $z_i\to\bar{z}(t)$ 的收敛速度(收敛阶).

定理 11.2 设定理 11.1 的条件成立,并且存在 h_0 和正常数 A_y, A_z, A_{y1}, A_{z1} 使得当 $h_y\leqslant h_0$, $h_z\leqslant h_0$ 时,成立估计式

$$\begin{aligned}\bar{e}_{\psi f}(h_y)\leqslant A_y h_y^{p_y},\quad e_{\text{inty}}\leqslant A_{y1}h_y^{q_y}\\ \bar{e}_{\psi g}(h_z)\leqslant A_z h_z^{p_z},\quad e_{\text{intz}}\leqslant A_{z1}h_z^{q_z}\end{aligned}\tag{11.51}$$

其中 p_y, p_z, q_y, q_z 是自然数, 则存在正常数A和 $\bar{h}_0$, 使得当 $h_y<\bar{h}_0$ 时,有估计式

$$\begin{cases}\|\bar{y}(t_i)-y_i\|\leqslant Ah_y^p,\\ \|\bar{z}(t_{ij})-z_{ij}\|\leqslant Ah_y^p,\end{cases}\tag{11.52}$$

其中 $p=\min\{p_y,\ p_z,\ q_y,\ q_z\}$.

证明　由估计式 (11.50) 和B的定义,将 (11.51) 代入, 即得所要的结果.

§4　稳定性分析

本节着重分析由梯形公式构成的组合方法的稳定性，其他的

线性方法构成的组合方法可以类似地进行分析。

考虑线性常系数系统

$$x' = Sx, \quad x(0) = x_0.$$

假设按慢变和快变分量分开，则可以写成形式为

$$y' = Ay + Bz, \quad y(0) = y_0, \tag{11.53}$$

$$z' = Cy + Dz, \quad z(0) = z_0, \tag{11.54}$$

其中 A、B 为M阶矩阵，C、D 为N阶矩阵，且

$$S = \begin{pmatrix} A & B \\ C & D \end{pmatrix}, \qquad x = \begin{pmatrix} y \\ z \end{pmatrix}, \qquad x' = \begin{pmatrix} y' \\ z' \end{pmatrix}.$$

我们用步长 $H = rh$，使用梯形公式先积分慢变子系统 (11.53) 一步；再用步长 h 的梯形公式积分快变子系统 (11.54) r 步. 对 (11.53) 用步长H积分时有

$$\begin{aligned} y_{mr} &= y_{(m-1)r} + \frac{H}{2}(y'_{mr} + y'_{(m-1)r}) \\ &= y_{(m-1)r} + \frac{H}{2}(Ay_{mr} + Bz_{mr} + Ay_{(m-1)r} + Bz_{(m-1)r}), \end{aligned} \tag{11.55}$$

这里，对 z_{mr} 量采用插值公式

$$\begin{aligned} z_{mr} &= z_{(m-1)r} + Hz'_{(m-1)r} \\ &= z_{(m-1)r} + HCy_{(m-1)r} + HDz_{(m-1)r} \end{aligned} \tag{11.56}$$

来计算. 将其代入 (11.55) 并整理得到

$$\begin{aligned} y_{mr} = &\left[I - \frac{H}{2}A\right]^{-1}\left[I + \frac{H}{2}A + HBC\right]y_{(m-1)r} \\ &+ \left[I - \frac{H}{2}A\right]^{-1}\left[HB + \frac{H^2BD}{2}\right]z_{(m-1)r}. \end{aligned}$$

令

$$\begin{aligned} \alpha_1 &= \left[I - \frac{H}{2}A\right]^{-1}\left[I + \frac{H}{2}A + HBC\right], \\ \beta_1 &= \left[I - \frac{H}{2}A\right]^{-1}\left[HB + \frac{H^2BD}{2}\right], \end{aligned} \tag{11.56'}$$

则上式为

$$y_{mr}=\alpha_1 y_{(m-1)r}+\beta_1 z_{(m-1)r}. \tag{11.57}$$

对 (11.54) 用步长 h 积分一步时，有

$$z_{n+1}=z_n+\frac{h}{2}(Cy_n+Dz_n+Cy_{n+1}+Dz_{n+1}),$$

整理得到

$$z_{n+1}=\left[I-\frac{h}{2}D\right]^{-1}\frac{hC}{2}y_n+\left[I-\frac{h}{2}D\right]^{-1}$$

$$\cdot\left[I+\frac{h}{2}D\right]z_n+\left[I-\frac{h}{2}D\right]^{-1}\frac{hC}{2}y_{n+1}$$

对 y_{n+1} 量采用插值公式

$$y_{n+1}=y_n+hy'_n=y_n+h(Ay_n+Bz_n)$$

计算，并令

$$\alpha_2=\left[I-\frac{h}{2}D\right]^{-1}\frac{hC}{2},$$

$$\beta_2=\left[I-\frac{h}{2}D\right]^{-1}\left[I+\frac{h}{2}D\right],$$

$$\gamma=\left[I-\frac{h}{2}D\right]^{-1}\frac{hC}{2}, \tag{11.57'}$$

则上式变为

$$z_{n+1}=\alpha_2 y_n+\beta_2 z_n+\gamma(y_n+hAy_n+hBz_n)$$

$$=(\alpha_2+\gamma+\gamma hA)y_n+(\beta_2+\gamma hB)z_n. \tag{11.58}$$

对 (11.54) 用步长 h 积分二步时，有

$$z_{n+2}=\alpha_2 y_{n+1}+\beta_2 z_{n+1}+\gamma y_{n+2},$$

并采用插值公式

$$y_{n+2}=y_n+2h(Ay_n+Bz_n)$$

计算 y_{n+2} 的值，则有

$$z_{n+2}=\left(\sum_{k=1}^{2}\beta_2^{2-k}\alpha_2+\sum_{k=1}^{1}(\beta_2^{1-k}\alpha_2)khA\right.$$

$$\left.+\sum_{k=1}^{2}\beta_2^{2-k}\gamma+\sum_{k=1}^{2}(\beta_2^{2-k}\gamma)khA\right)y_n$$

$$+\left(\beta_2^2+\sum_{k=1}^{1}(\beta_2^{1-k}\alpha_2)khB\right.$$

$$\left.+\sum_{k=1}^{2}(\beta_2^{2-k}\gamma)khB\right)z_n. \tag{11.59}$$

对 (11.54) 用步长 h 积分三步时，有

$$z_{n+3}=\alpha_2 y_{n+2}+\beta_2 z_{n+2}+\gamma y_{n+3}$$

并采用 $y_{n+3}=y_n+3h(Ay_n+Bz_n)$ 来计算 y_{n+3} 的值，则有

$$z_{n+3}=\left(\sum_{k=1}^{3}\beta_2^{3-k}\alpha_2+\sum_{k=1}^{2}(\beta_2^{2-k}\alpha_2)khA\right.$$

$$\left.+\sum_{k=1}^{3}(\beta_2^{3-k}\gamma)+\sum_{k=1}^{3}(\beta_2^{3-k}\gamma)khA\right)y_n$$

$$+\left(\beta_2^3+\sum_{k=1}^{2}(\beta_2^{2-k}\alpha_2)khB\right.$$

$$\left.+\sum_{k=1}^{3}(\beta_2^{3-k}\gamma)khB\right)z_n. \tag{11.60}$$

对 (11.54) 用步长 h 积分 r 步时，有

$$z_{n+r}=\alpha_2 y_{n+r-1}+\beta_2 z_{n+r-1}+\gamma y_{n+r}.$$

同样，采用 $y_{n+r}=y_n+rh(Ay_n+Bz_n)$ 来计算 y_{n+r} 的值，则有

$$z_{n+r}=\left(\sum_{k=1}^{r}\beta_2^{r-k}\alpha_2+\sum_{k=1}^{r-1}(\beta_2^{r-1-k}\alpha_2)khA\right.$$

$$\left.+\sum_{k=1}^{r}\beta_2^{r-k}\gamma+\sum_{k=1}^{r}(\beta_2^{r-k}\gamma)khA\right)y_n$$

$$+\left(\beta_2^r+\sum_{k=1}^{r-1}(\beta_2^{r-1-k}\alpha_2)khB\right.$$

$$\left.+\sum_{k=1}^{r}(B_2^{r-k}\gamma)khB\right)z_n. \tag{11.61}$$

令

$$\alpha_2^*=\left(\sum_{k=1}^{r}\beta_2^{r-k}\alpha_2+\sum_{k=1}^{r-1}(\beta_2^{r-1-k}\alpha_2)khA\right.$$

$$+\sum_{k=1}^{r}\beta_2^{r-k}\gamma+\sum_{k=1}^{r}(\beta_2^{r-k}\gamma)khA\Big),$$

$$\beta_2^*=\Big(\beta_2^r+\sum_{k=1}^{r-1}(\beta_2^{r-1-k}\alpha_2)khB \tag{11.61'}$$

$$+\sum_{k=1}^{r}(\beta_2^{r-k}\gamma)khB\Big),$$

则 (11.61) 变成为

$$z_{n+r}=\alpha_2^*y_n+\beta_2^*z_n \tag{11.62}$$

若令 $n=(m-1)r$，且把 (11.57) 与 (11.62) 两式结合起来，则可写成矩阵形式

$$\begin{pmatrix} y_{mr} \\ z_{mr} \end{pmatrix}=\begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2^* & \beta_2^* \end{pmatrix}\begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}. \tag{11.63}$$

定义 11.1 如果将上式梯形公式构成的组合方法应用于方程组 (11.53) 和 (11.54)得公式 (11.63)，当 $m\to\infty$ 时，$y_{mr}\to 0$，$z_{mr}\to 0$，则称该方法为绝对稳定的.

显然有如下的定理成立:

定理 11.3 上述梯形公式构成的组合方法的绝对稳定的充要条件是矩阵

$$\begin{pmatrix} \alpha_1 & \beta_1 \\ \alpha_2^* & \beta_2^* \end{pmatrix}$$

的特征值的模小于 1 .

这个定理应用起来需要许多计算，下面讨论更加简单的情形.

对 (11.53) 用步长H积分时，如果不考虑对所需要的另一个子系统的解分量进行插值，且假定 $z_{mr}=z_{(m-1)r}$，则 (11.55) 为

$$y_{mr}=\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right]y_{(m-1)r}$$

$$+\left[I-\frac{H}{2}A\right]^{-1}HBz_{(m-1)r}.$$

(11.56′) 式中的 α_1 和 β_1 变为

$$\alpha_1=\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right],$$

$$\beta_1 = \left[I - \frac{H}{2}A\right]^{-1} HB,$$

所以有

$$y_{mr} = \alpha_1 y_{(m-1)r} + \beta_1 z_{(m-1)r}. \tag{11.64}$$

这个算法步骤写成矩阵形式为

$$\begin{pmatrix} y_{mr} \\ z_{(m-1)r} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}, \tag{11.65}$$

对 (11.54) 从 $t_n(n=(m-1)r)$ 积分一个步长 h 时，同样如果不考虑对所需要的另一个分量的值进行插值，且假定 $y_{mr} = y_{(m-1)r}$，则有

$$z_{n+1} = \left[I - \frac{h}{2}D\right]^{-1} hCy_{mr} + \left[I - \frac{h}{2}D\right]^{-1}\left[I + \frac{h}{2}D\right] z_n, \tag{11.66}$$

即

$$z_{n+1} = \alpha_2 y_{mr} + \beta_2 z_n,$$

这里的 α_2, β_2 分别为

$$\alpha_2 = \left[I - \frac{hD}{2}\right]^{-1} hC, \quad \beta_2 = \left[I - \frac{hD}{2}\right]^{-1}\left[I + \frac{hD}{2}\right].$$

同样，这个算法步骤也可以写为矩阵形式

$$\begin{pmatrix} y_{mr} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \alpha_2 & \beta_2 \end{pmatrix} \begin{pmatrix} y_{mr} \\ z_n \end{pmatrix}. \tag{11.67}$$

定义 11.2 若有

$$P = \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & I \end{pmatrix}, \qquad Q = \begin{pmatrix} I & 0 \\ \alpha_2 & \beta_2 \end{pmatrix}$$

称 $R = Q^r P$ 为压缩矩阵.

定理 11.4 上述梯形公式构成的组合方法(不考虑对所需要的另一个子系统的解分量进行插值的情形)的绝对稳定的充要条件是压缩矩阵 R 的特征值的模小于 1.

证明由对 (11.53) 用步长 H 积分一步，有

$$\begin{pmatrix} y_{mr} \\ z_{(m-1)r} \end{pmatrix} = \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}.$$

由对 (11.54) 用步长 h 积分一步，有

$$\begin{pmatrix} y_{mr} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \alpha_2 & \beta_2 \end{pmatrix} \begin{pmatrix} y_{mr} \\ z_{(m-1)r} \end{pmatrix}.$$

若积分 r 步，则有

$$\begin{pmatrix} y_{mr} \\ z_{n+r} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \alpha_2 & \beta_2 \end{pmatrix}^r \begin{pmatrix} y_{mr} \\ z_n \end{pmatrix}.$$

令 $n=(m-1)r$，则得

$$\begin{pmatrix} y_{mr} \\ z_{mr} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \alpha_2 & \beta_2 \end{pmatrix}^r \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}$$

$$= \begin{pmatrix} I & 0 \\ \sum_{j=1}^{r} \left[I - \frac{h}{2}D\right]^{-j} \left[I + \frac{h}{2}D\right]^{j-1} hC & \left[I - \frac{h}{2}D\right]^{-r} \left[I + \frac{h}{2}D\right]^{r} \end{pmatrix}$$

$$\cdot \begin{pmatrix} \alpha_1 & \beta_1 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}.$$

由定义 11.2，我们有

$$\begin{pmatrix} y_{mr} \\ z_{mr} \end{pmatrix} = Q^r P \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix} = R \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix}$$

即得出该方法绝对稳定的充要条件为 $R=Q^rP$ 的特征值的模小于1．定理证毕．

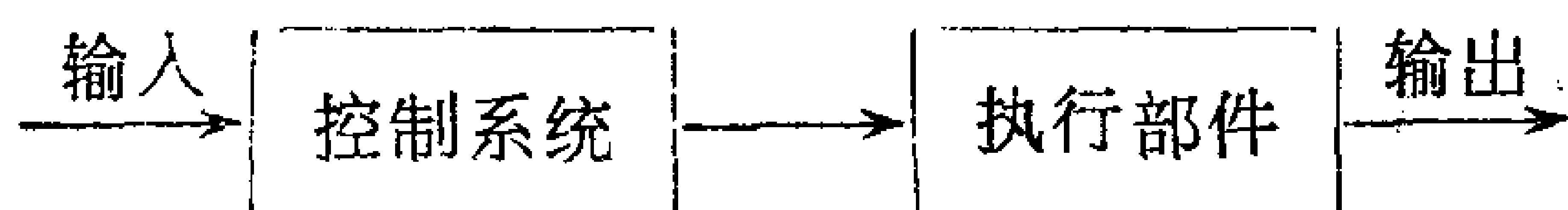
由上述定理，我们还可得到该方法绝对稳定的另一种充分条件，即：

定理 11.5 若 $\|P\|<1$ 和 $\|Q\|<1$，则该方法是绝对稳定的．

由 P、Q 的定义，我们知道这个条件与给定的系统有关，可以直接进行验算．

定义 11.3 若(11.53)式中的矩阵 $B=0$，或(11.54)式中的矩阵 $C=0$，即矩阵 S 为块三角形矩阵，则称该系统为弱影响系统．

开环控制系统，一般都可用弱影响系统的形式来表示．例如以下开环系统



控制系统的变量用 z 表示，执行部件的变量用 y 表示．控制系统的运动不受执行部件的运动状态的影响．但是控制系统的运动状态则影响执行部件的运动．在这种情形下有如下定理：

定理 11.6 如果 S 是块三角形矩阵（$B=0$ 或 $C=0$），即系统为弱影响系统，那么由上述梯形公式构成的组合方法的绝对稳定的充要条件是 A 和 D 的特征值的实部小于零．

证明 如果 $B=0$．先讨论本节开始研究的带插值的梯形公式构成的组合方法的稳定性．由 (11.56′) 式知道

$$\alpha_1=\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right],\quad \beta_1=0.$$

由 (11.61′) 式知道

$$\beta_2^*=\beta_2^r=\left[I-\frac{h}{2}D\right]^{-r}\left[I+\frac{h}{2}D\right]^r.$$

这样，(11.63) 就成为

$$\begin{pmatrix} y_{mr} \\ z_{mr} \end{pmatrix}=\begin{pmatrix} \alpha_1 & 0 \\ \alpha_2^* & \beta_2^r \end{pmatrix}\begin{pmatrix} y_{mr} \\ z_{(m-1)r} \end{pmatrix}.$$

我们知道块三角形矩阵

$$\begin{pmatrix} \alpha_1 & 0 \\ \alpha_2^* & \beta_2^r \end{pmatrix}=\begin{pmatrix} \left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right] & 0 \\ \alpha_2^* & \left[I-\frac{h}{2}D\right]^{-r}\left[I+\frac{h}{2}D\right]^r \end{pmatrix}$$

的特征值就是块对角矩阵的特征值．其特征值的模小于 1，当且仅当 A 和 D 的特征值的实部小于零．

对于不带插值的梯形公式构成的组合方法，即简单情形的上述组合方法，由定义 11.2，有

$$P=\begin{pmatrix}\alpha_1 & 0\\ 0 & I\end{pmatrix}$$

$$=\begin{pmatrix}\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right] & 0\\ 0 & I\end{pmatrix}$$

$$Q=\begin{pmatrix}I & 0\\ \alpha_2 & \beta_2\end{pmatrix}$$

$$=\begin{pmatrix}I & 0\\ \left[I-\frac{h}{2}D\right]^{-1}hC & \left[I-\frac{h}{2}D\right]^{-1}\left[I+\frac{h}{2}D\right]\end{pmatrix}.$$

所以,

$$R=Q^rP=\begin{pmatrix}I & 0\\ \alpha_2 & \beta_2\end{pmatrix}\begin{pmatrix}\alpha_1 & 0\\ 0 & I\end{pmatrix}$$

$$=\begin{pmatrix}I & 0\\ P & \left[I-\frac{h}{2}D\right]^{-r}\left[I+\frac{h}{2}D\right]^{r}\end{pmatrix}$$

$$\cdot\begin{pmatrix}\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right] & 0\\ 0 & I\end{pmatrix}$$

$$=\begin{pmatrix}\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right] & 0\\ P\left[I-\frac{H}{2}A\right]^{-1}\left[I+\frac{H}{2}A\right] & \left[I-\frac{h}{2}D\right]^{-r}\left[I+\frac{h}{2}D\right]^{r}\end{pmatrix},$$

其中

$$P=\sum_{j=1}^{r}\left[I-\frac{h}{2}D\right]^{-j}\left[I+\frac{h}{2}D\right]^{j-1}hC.$$

同样，压缩矩阵R的特征值就是块对角矩阵的特征值．因此其特征值的模小于1，当且仅当A和D的特征值的实部小于零．

若$C=0$时，类同上面的证明，可得到同样的结果．

这个定理说明，用梯形公式构成的组合方法应用于弱影响系统，仍然保持其A稳定性的性质．

定理 11.7 如果 S 是一个块三角形矩阵（若 $B=0$），即系统是弱影响系统，算法 II 构成的组合方法绝对稳定的充要条件是 A 的特征值位于隐式 Runge-Kutta 方法的绝对稳定区域内，D 的特征值位于显式 Runge-Kutta 方法的绝对稳定区域内.

证明 将算法II应用于方程组

$$z' = Az, \tag{11.68}$$

$$y' = Cz + Dy, \tag{11.69}$$

这里

$$S = \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}.$$

对 (11.69) 式，用步长 H 使用显式 Runge-Kutta 方法积分一步，有

$$\begin{pmatrix} z_{(m-1)r} \\ y_{mr} \end{pmatrix} = \begin{pmatrix} I & 0 \\ Q & \left[I + HD + \frac{1}{2!}(HD)^2 + \frac{1}{3!}(HD)^3 + \frac{1}{4!}(HD)^4\right] \end{pmatrix} \cdot \begin{pmatrix} y_{(m-1)r} \\ z_{(m-1)r} \end{pmatrix},$$

其中

$$Q = \left[7/6 + \frac{5}{6}(HD) + \frac{1}{6}(HD)^2 + \frac{1}{12}(HD)^3\right](HC).$$

对 (11.68) 式，用步长 h，使用隐式 Runge-Kutta 方法积分 r 步，有

$$\begin{pmatrix} z_{mr} \\ y_{mr} \end{pmatrix} = \begin{pmatrix} \left[I - \frac{h}{2}A + \frac{(hA)^2}{12}\right]^{-1}\left[I + \frac{h}{2}A + \frac{(hA)^2}{12}\right] & 0 \\ 0 & I \end{pmatrix}^r$$

$$\cdot \begin{pmatrix} z_{(m-1)r} \\ y_{mr} \end{pmatrix}$$

于是,得到

$$\begin{pmatrix} z_{mr} \\ y_{mr} \end{pmatrix} = \begin{pmatrix} \left[I - \frac{h}{2}A + \frac{(hA)^2}{12}\right]^{-r}\left[I + \frac{h}{2}A + \frac{(hA)^2}{12}\right]^r & 0 \\ 0 & I \end{pmatrix}$$

$$\cdot \begin{pmatrix} I & 0 \\ Q\left[I + HD + \frac{1}{2!}(HD)^2 + \frac{1}{3!}(HD)^3 + \frac{1}{4!}(HD)^4\right] \end{pmatrix}$$

$$\cdot \begin{pmatrix} z_{(m-1)r} \\ y_{(m-1)r} \end{pmatrix},$$

$$R = Q'P = \begin{pmatrix} \left[I - \frac{hA}{2} + \frac{(hA)^2}{12}\right]^{-r}\left[I + \frac{hA}{2} + \frac{(hA)^2}{12}\right]^r & 0 \\ Q & \left[I + HD + \frac{1}{2!}(HD)^2 + \frac{1}{3!}(HD)^3 + \frac{1}{4!}(HD)^4\right] \end{pmatrix}.$$

同上,R 的特征值也就是块对角矩阵的特征值. 因此,其特征值的模小于 1 ,当且仅当 A 的特征值在隐式 Runge-Kutta 方法的绝对稳定区域内, D 的特征值在显式 Runge-Kutta 方法的绝对稳定区域之内. 所以,定理得证.

这样详细地写出定理的证明,目的是使读者进一步熟悉组合方法及其稳定性分析的基本思想. 其实对弱影响系统而言,组合方法的数值稳定性结论是十分显然的. 例如有弱影响系统

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & D \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}$$

先求解方程组 $x' = Ax$, 数值方法的稳定性依赖于矩阵 A 的特征值是否位于方法的绝对稳定区域内. 然后求解 $y' = Cx + Dy$. 这时对 x 进行插值,方法的稳定性由矩阵 D 的特征值来决定,看其是

否位于方法的绝对稳定的区域内．因此，在这种情形下，组合方法的数值稳定性条件与原来的未组合的数值积分方法的稳定性条件相同．

在 S 是块三角矩阵的情形，即系统为弱影响系统，系数矩阵的特征值可以用来描述组合方法的绝对稳定性．但是，对于一般的系统而言，S 的特征值却不能确定 R 的特征值．例如，我们取

$$S_1=\begin{pmatrix}-3 & 4\\ -2 & 2\end{pmatrix},\quad S_2=\begin{pmatrix}-1 & -2\\ 1 & 0\end{pmatrix},$$

这里 S_1 和 S_2 是相似矩阵，它们都有特征值 $\dfrac{-1\pm\sqrt{7}\,i}{2}$．我们用步长 $H=h=1$，采用向后 Euler 方法构成组合方法，求解

$$\begin{pmatrix}y'\\ z'\end{pmatrix}=S_1\begin{pmatrix}y\\ z\end{pmatrix},$$

由于 $R_1=QP=\begin{pmatrix}1 & 0\\ 2 & -1\end{pmatrix}\begin{pmatrix}\frac{1}{4} & 1\\ 0 & 1\end{pmatrix}=\begin{pmatrix}\frac{1}{4} & 1\\ \frac{1}{2} & 1\end{pmatrix}$，它有特征值 $\dfrac{5\pm\sqrt{41}}{8}$，其中有一个特征值的模大于 1．因此，在某个时刻开始解是发散的．

若用同样的方法求解相似系统

$$\begin{pmatrix}y'\\ z'\end{pmatrix}=S_2\begin{pmatrix}y\\ z\end{pmatrix},$$

则有

$$R_2=QP=\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\begin{pmatrix}\frac{1}{2} & -1\\ 0 & 1\end{pmatrix}=\begin{pmatrix}\frac{1}{2} & -1\\ \frac{1}{2} & 0\end{pmatrix}$$

它有特征值 $\dfrac{1\pm\sqrt{7}\,i}{4}$，其模均小于 1，那么解总是趋于零的．所

以相似变换不能保持绝对稳定性。

本 章 附 注

本章和下一章的主要内容是作者自己在有关方面的经验总结。它是根据作者在实际研究工作中的部分论文报告和一些具体的处理思想写成的。

第十二章 自动控制系统常微分方程组的数值解法

自动控制系统的运动一般用常微分方程组的初值问题来描写，是一类典型的刚性初值问题．本章针对其特点讨论数值积分中的某些问题．

§1 问题的提出

自动控制系统运动过程，一般用常微分方程(或方程组)的初值问题来描写．自动控制系统设计有很多方法．对于复杂的，特别是非线性系统的设计，利用数值积分，求出系统的数值解，由此选择系统方案和参数，也是一个重要手段．对已设计好的系统，也可以用数值解检验系统的品质．因此，常微分方程初值问题的数值解法在控制系统的设计中起着重要的作用．

描写自动控制系统运动的常微分方程组，都包含快变分量和慢变分量，如控制部件主要由电子线路实现，是快变的，而执行部件和控制对象由机械实现，是慢变的，所以是典型的刚性初值问题．快变分量一般是指数衰减或指数振荡衰减，在很短的时间内完成，这个过程叫暂态过程或过渡过程．

从数学观点看，微分方程组有一个稳态解，稳态解与微分方程组的初值无关．初值所决定的微分方程组的解会很快地趋于稳态解，亦即两个解的差别很快地趋于零，这就是暂态过程．另一种情形是微分方程组的右端项有变化，或间断性的变化，这时微分方程组有了新的稳态解，这时方程组的解由原来的函数很快地变到新的稳态解，也是暂态过程．

多数的计算问题是暂态过程只出现一次，考察控制系统品质

的问题就是这样．在实际问题中一般是暂态过程多次出现，甚至一个暂态过程没有结束另一个暂态过程又出现．例如有人干预的过程暂态过程就会多次出现．

自动控制系统常用的描述方法是用每个环节的方程的 Laplace 变换，右端项的变换叫输入，函数的变换叫输出，算子叫做传递函数，用图 12.1 的形式表示．例如方程

$$ay'' + by' + y = cx' + x,$$

在这个方程中 $x(t)$ 看成已知函数，$y(t)$ 是方程的解．在自动控制系统的表示法是

$$Y(s) = W(s)X(s), \quad W(s) = \frac{cs+1}{as^2+bs+1}$$

所要指出的是，在变换中采用初值 $y(0) = \dot{y}(0) = 0$，$x(0) = 0$ 的形式(下同)，但在计算时初值并不一定是零，需要另外给出来．

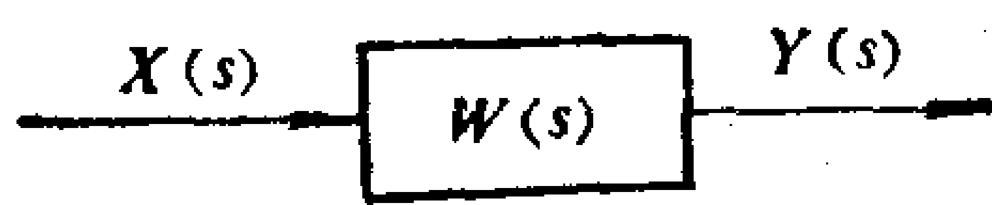


图 12.1 自动控制系统的环节

将输出和输入联接起来，就得到了描写整个系统的方框图，如图 12.2 中 $X(s)$ 是输入，$Y(s)$ 是输出，$Z(s)$，$V(s)$ 是中间函数，$V(s)$ 是反馈信号，$W_1(s)$ 框的输入是 $X(s) - V(s)$.

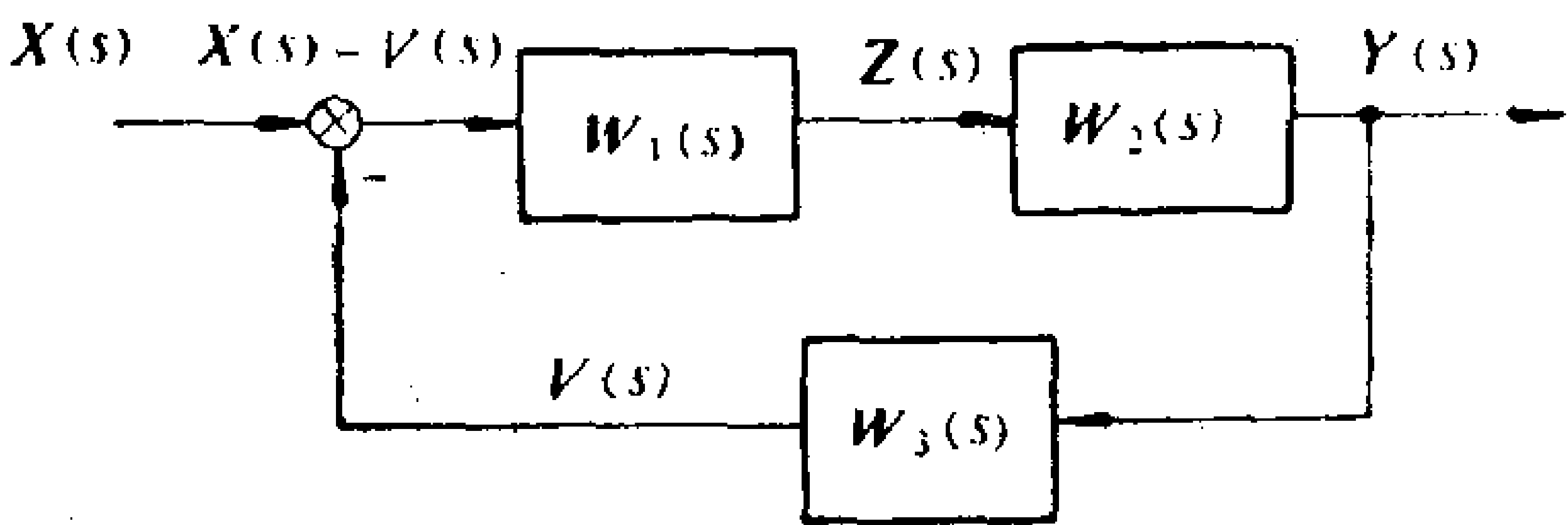


图 12.2 自动控制系统的方框图

控制部件中很多环节是由电阻、电容和电感做成的，描述这样的环节的方程是线性常系数的方程．经过 Laplace 变换，只要初值是零，不过是把小写字母表示的函数符号改成大写，把自变量换

成 s，把对 t 的微分符号 $\frac{d}{dt}$ 换成 s，就得到了变换后的方程.

例如图 12.3 所示的环节，设 $x(t)$ 和 $y(t)$ 分别是输入和输出电压，则其方程是

$$LCy'' + RCy' + y = x \tag{12.1}$$

经过 Laplace 变换后的方程是

$$(LCs^2 + RCs + 1)Y = X,$$

可写成

$$Y = \frac{1}{LCs^2 + RCs + 1} X \tag{12.2}$$

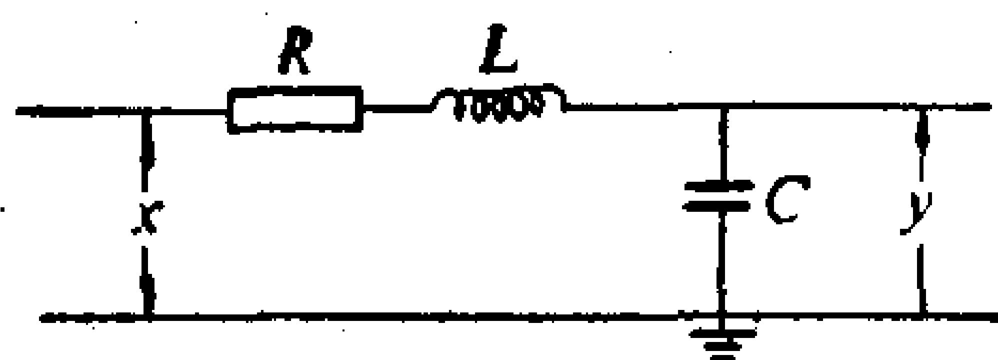


图 12.3 电阻电感电容组成的环节

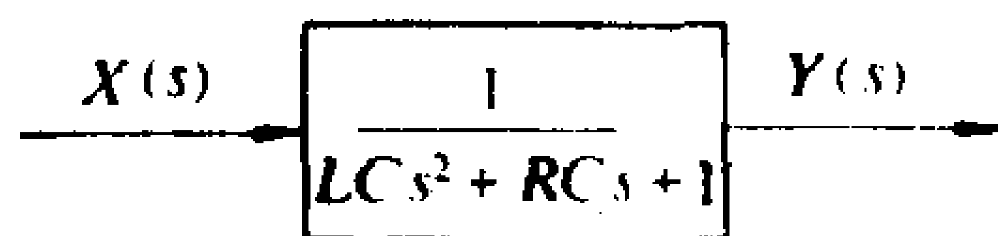


图 12.4 环节(12.1)的方框图

其方框图的形式是图 12.4.

今后我们不再区分 (12.1) 和 (12.2)，即不再区分函数 $x(t)$ 和其 Laplace 变换 $X(s)$，只是把 s 做为算子符号 $\frac{d}{dt}$ 来理解.所要注意的是把图 12.5 所示的时延环节理解为

$$y(t) = x(t - \tau)$$



图 12.5 时延环节

对于非线性环节，一般是在方框中直接指出其函数关系．如图 12.6 表示开关电路，其函数关系是

$$y = \begin{cases} 1 & x \geq 0, \\ -1 & x < 0. \end{cases}$$

又如图 12.7 表示限幅器电路，其函数关系是

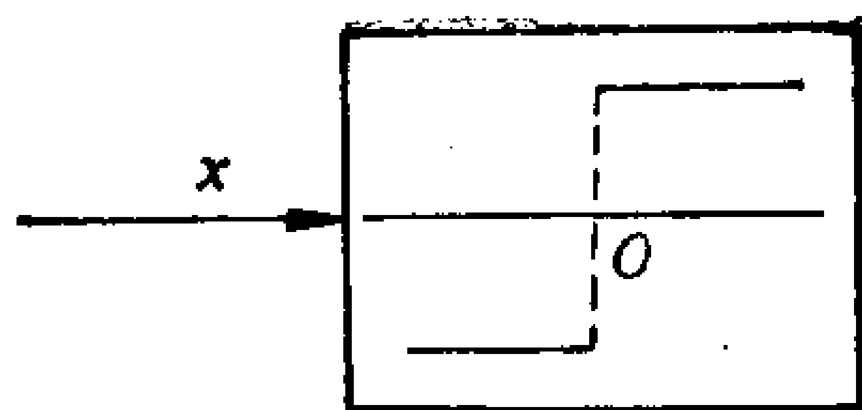


图 12.6　开关电路

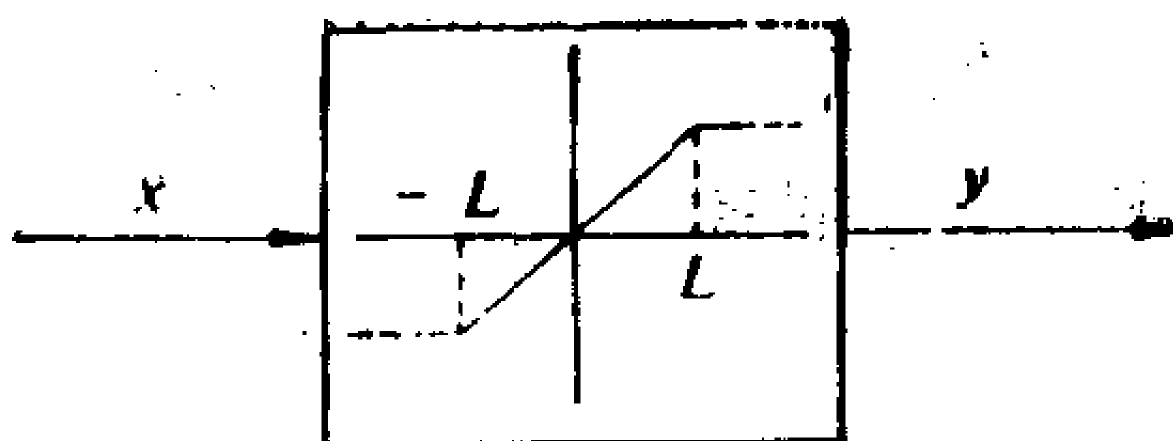


图 12.7　限幅器电路

$$y = \begin{cases} x & |x| \leqslant L, \\ L\operatorname{sign}x & |x| > L. \end{cases}$$

在图 12.6 和图 12.7 中 $x(t)$ 是输入，$y(t)$ 是输出.

从自动控制系统的方框图很容易写出它的常微分方程组的形式. 这种方程组的刚性性质表现在，若用有限的稳定区域的数值方法积分时，快变分量决定着要用很小的步长，这在暂态过程是必要的，但是在暂态过程结束以后，快变分量已经小得可以忽略，本来从慢变分量的角度看，可以用较大的步长积分，但是理论和实践都说明，步长仍要由快变分量决定，即步长不能增大，否则误差会指数增长，即计算不稳定.

我们举例说明这个问题[19]. 例如对一阶放大器环节，如图 12.8 所示.

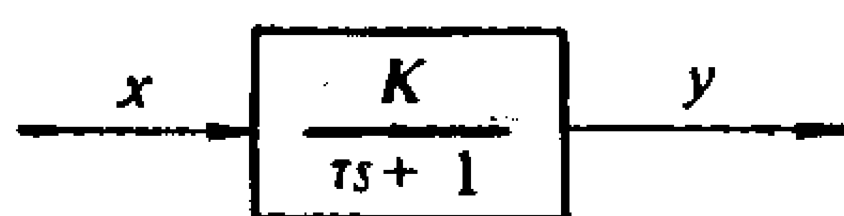


图 12.8　放大器环节

其方程为

$$\tau y' + y = Kx. \tag{12.3}$$

设方程是孤立的，$x(t)$ 是阶跃函数

$$x(t) = \begin{cases} 0 & t < 0, \\ 1 & t \geqslant 0, \end{cases}$$

$y(0) = 0$. 设要从 $t = 0$ 做数值积分. 首先做变换

$$u(t) = K - y(t)$$

则在 $t \geqslant 0$，方程变成

$$\tau u' + u = 0,\ u(0) = K. \tag{12.4}$$

这个方程的解是 $u(t) = Ke^{-t/\tau}$ 或 $y(t) = K(1 - e^{-t/\tau})$。$\tau$ 是很小的正数，$u(t) \to 0$，$y(t) \to K$ 的速度很快。

若用 Euler 公式解后一问题，则得

$$u_n = K(1 - h/\tau)^n,$$

这里 h 是积分步长，n 是积分步数，$t = nh$，u_n 是 $u(t)$ 的近似值。由此可见：当 $n \to +\infty$ 时

(1) 若 $h < 2\tau$，则 $u_n \to 0$；

(2) 若 $h > 2\tau$，则 $|u_n| \to +\infty$；

(3) 若 $h = 2\tau$，则 $u_n = (-1)^n K$。

图 12.9 给出了真解 $e^{-t/\tau}$ 和不同步长 h 的数值解。

因为微分方程和数值方法都是线性的，所以初值的误差仍满

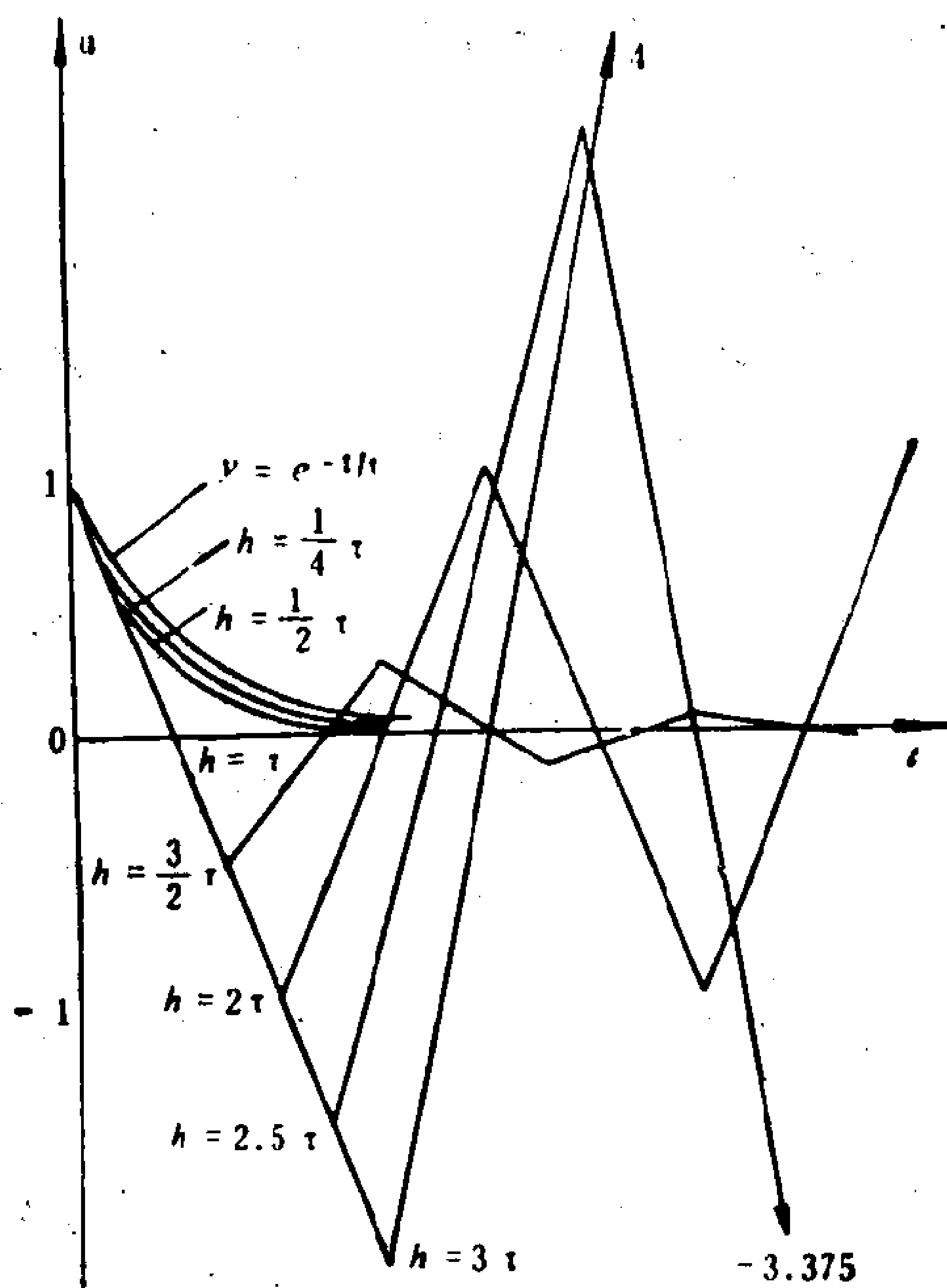


图 12.9　真解 $e^{-t/\tau}$ 和 u_n 的图形

足同一个方程和数值积分公式. 因而设 $u(0)$ 有误差 $\triangle u_0$, 则微分方程解和数值解的误差分别是

$$\triangle u = \triangle u_0 e^{-t/\tau}, \quad \triangle u_n = \triangle u_0 (1 - h/\tau)^n.$$

由计算稳定性的定义可知 $h < 2\tau$ 时 Euler 法稳定, $h \geq 2\tau$ 时不稳定. 由于 τ 很小,所以要取很小的步长. 我们注意,假如这个方程是方程组中的一个方程,当 u_n 小到可以忽略时,步长 h 也不能放大,因为若要 $h > 2\tau$, 这个方程的误差,就象初值 $\triangle u_0$ 引起的误差那样,很快地增长到不可思议的程度.

现在我们分析最简单的复特征根的情形. 设方程和初值分别是

$$z' = \lambda z, \quad z(0) = 1, \tag{12.5}$$

这里 $\lambda = \alpha + i\beta$, $\alpha < 0$. 方程的解是

$$z(t) = e^{\alpha t}(\cos\beta t + i\sin\beta t).$$

令 $z = x + iy$, 分离实部虚部,得方程组

$$x' = \alpha x - \beta y, \quad x(0) = 1,$$
$$y' = \beta x + \alpha y, \quad y(0) = 0,$$

消去 y, 得

$$x'' - 2\alpha x' + (\alpha^2 + \beta^2)x = 0, \quad x(0) = 1, \quad x'(0) = \alpha.$$

y 满足同一个微分方程,但初值是

$$y(0) = 0, \quad y'(0) = \beta.$$

用 Euler 法解 (12.5) 时,当 $|1 + h\lambda| < 1$ 时 $z_n \to 0$, 或初值误差 $\triangle z_0$ 引起的误差 $\triangle z_n \to 0$, 这个条件等价于

$$h < 2(-\alpha)/(\alpha^2 + \beta^2).$$

由 (12.5) 的解可知, $|\alpha|$ 越大,解衰减越快; β 越大, 解的振荡越快. 由此条件可知:

(1) 若 $-\alpha$ 和 β 成比例地增大, h 要减小,即解的衰减和振荡都加快,步长要减小.

(2) 若 β 不变,而 $-\alpha$ 增大, 当 $-\alpha < \beta$ 时, h 可以增大, 当 $-\alpha > \beta$ 时, h 要减小. 也就是说振荡频率不变而衰减加快时,在衰减较小时(振荡起主要作用时)步长可以增加, 但在衰减较大

时(衰减起主要作用时),步长要缩小.

(3) 若$-\alpha$不变,而β增大时, h要缩小. 即衰减不变, 振荡频率增大时,步长要缩小.

和上边的讨论一样, 若方程(12.5)是方程组的一个方程, 不管它的解多么小,数值积分步长都要受这个方程稳定条件的限制.

§2 计算稳定性

上节我们讨论了用 Euler 法解方程(12.4), (12.5)时的稳定问题. 研究如此简单的微分方程数值解的计算稳定问题, 原因在于:

(1) 这种方程有很广泛的代表性, 线性常系数常微分方程组通过线性变换可变到这种情形,或类似的情形,所以我们把微分方程 $y'=\lambda y(\mathrm{Re}\lambda<0)$ 叫做试验方程;

(2) 线性非齐次微分方程组数值解的计算稳定性和齐次方程组是一样的,所以只研究齐次方程组的计算稳定性就够了;

(3) 变系数的常微分方程组, 在工程上常常将系数固态化后研究某个时刻邻域的性质,这化成了研究常系数的微分方程组.我们知道固态化后的问题与原来变系数的问题并不完全等价, 但在很多情况下是可用的.

(4) 非线性的常微分方程组,在研究某个解邻域的误差时,可以经过线性化,得到误差所满足的线性方程组,对数值解也有同样的情形.

现在我们讨论试验方程的代表性的问题. 对于线性常系数齐次常微分方程组

$$y'=Ay, \tag{12.6}$$

这里 $y\in R^m$, A是m阶矩阵.

若A的 Jordan 标准形的 Jordan 块都是一维的,即存在非奇异矩阵 S, 使

$$S^{-1}AS=D=\mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m).$$

令 $z = S^{-1}y$，则 (12.6) 化为

$$z' = Dz,$$

即 (12.6) 经过线性变换，变成了 m 个互不相交的方程

$$z_i' = \lambda_i z_i, \quad i = 1, 2, \cdots, m,$$

每个方程都是试验方程的形式，这时讨论试验方程的计算稳定性，就代表了讨论方程组的计算稳定性.

若 A 的 Jordan 标准形中的 Jordan 块有大于一维的，经过线性变换可以把方程组 (12.6) 化成一些互不相干的组，此时无妨设 A 化成一个 m 维的 Jordan 块，即存在非奇异矩阵 S，使

$$S^{-1}AS = J = \begin{pmatrix} \lambda & & & \\ 1 & \lambda & & \\ & \ddots & \ddots & \\ & & 1 & \lambda \end{pmatrix}.$$

这时相应的齐次方程组是

$$\begin{aligned} z_1' &= \lambda z_1, \\ z_k' &= z_{k-1} + \lambda z_k, \quad k = 2, \cdots, m. \end{aligned} \tag{12.7}$$

若用 Euler 法解方程组 (12.7)，则有

$$z_{n+1} = Bz_n,$$

这里

$$B = (1 + h\lambda)E + hE_1,$$

E 为 m 阶单位矩阵，E_1 是主对角下为 1 其余元为零的 m 阶矩阵. 我们知道

$$E_1^n = 0, \text{ 当 } n \geqslant m,$$

而

$$B^n = \sum_{j=0}^{m-1} C_j^n (1 + h\lambda)^{n-j} h^j E_1^j$$

只要

$$|1 + h\lambda| < 1,$$

将有 $B^n \to 0 (n \to \infty)$.

这个条件和用 Euler 法解试验方程的计算稳定条件是一样的.

一般用隐式或显式单步法解微分方程组(12.7)，可得前后两步的如下关系

$$y_{n+1}=[\varphi(hJ)]^{-1}\psi(hJ)y_n,$$

这里 φ，ψ 是多项式．因为 $J=\lambda E+E_1$，可以看出

$$\varphi(hJ)=\varphi(h\lambda)E+F$$
$$\psi(hJ)=\psi(h\lambda)E+F$$

这里的 F 泛指主对角元素为零的 m 阶下三角矩阵．又因为 $\varphi(hJ)$ 是主对角元素为 $\varphi(h\lambda)$ 的下三角矩阵，所以

$$[\varphi(hJ)]^{-1}=[\varphi(h\lambda)]^{-1}E+F,$$

将两个矩阵相乘

$$B=[\varphi(hJ)]^{-1}\psi(hJ)=\frac{\psi(h\lambda)}{\varphi(h\lambda)}E+F,$$

又由 $F^n=0$，若 $n\geqslant m$，所以

$$B^n=\sum_{j=0}^{m-1}C_j^n\left[\frac{\psi(h\lambda)}{\varphi(h\lambda)}\right]^{n-j}F^j,$$

F^j 与 n 无关，C_j^n 是 n 的幂次，而 $\left[\frac{\psi(h\lambda)}{\varphi(h\lambda)}\right]^{n-j}$ 是某个数值的 n 次幂，所以若

$$\frac{\psi(h\lambda)}{\varphi(h\lambda)}<1,\tag{12.8}$$

当 $n\to\infty$ 时，$B^n\to 0$．

不等式(12.8)是将同一个数值方法用于试验方程的计算稳定条件．

从以上事实我们可以得出结论：

定理 12.1 设一阶线性常系数常微分方程组

$$\dot{y}=Ay+F(t)\tag{12.9}$$

系数矩阵 A 的特征值是 $\lambda_1,\lambda_2,\cdots,\lambda_m$．对于步长 $h>0$，假若一个单步数值积分法用于试验方程 $\dot{y}=\lambda_j y(j=1,\cdots,m)$ 计算稳定，那么这个单步法用于方程组(12.9)时也是计算稳定的．

以下我们讨论线性多步法的情形．令

$$\rho(s)=\sum_{l=0}^{k}\alpha_l s^l,\ \sigma(s)=\sum_{l=0}^{k}\beta_l s^l,\ E_{y_n}=y_{n+1},$$

考虑线性多步法

$$\rho(E)y_n=h\sigma(E)\dot{y}_n,$$

将这个方法用干齐次线性常系数常微分方程组

$$\dot{y}=Ay.$$

我们有

$$\rho(E)y_n=hA\sigma(E)y_n,$$

即是

$$\sum_{l=0}^{k}(\alpha_l I_m-h\beta_l A)y_{n+l}=0,$$

这里 I_m 是 m 阶单位矩阵. 为了能解出 y_{n+k}，应当假设 $\alpha_k I_m-h\beta_k A$ 的逆存在，实际上，若 $\alpha_k\neq 0$，只要取得 $h\neq\frac{\alpha_k}{\lambda\beta_k}$ (λ 是 A 的特征值)，逆矩阵就存在. 所以有

$$y_{n+k}=-(\alpha_k I_m-h\beta_k A)^{-1}\sum_{l=0}^{k-1}(\alpha_l I_m-h\beta_l A)y_{n+l}.$$

将它化为两层的迭代形式. 令

$$y_n^{(j)}=y_{n+j},\ j=0,1,\cdots,k-1,$$

上式可写戊

$$z_{n+1}=Bz_n$$

这里

$$z_n=\begin{pmatrix}y_n^{(0)}\\ y_n^{(1)}\\ \vdots\\ y_n^{(k-1)}\end{pmatrix},\quad B=\begin{pmatrix}0 & I_m & 0 & \cdots & 0\\ 0 & 0 & I_m & \cdots & 0\\ & & & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & I_m\\ * & * & * & \cdots & *\end{pmatrix},$$

B 的元素 I_m 为 m 阶矩阵的 k 阶矩阵，即 B 为 $m\times k$ 阶矩阵，z_n 是 $m\times k$ 维向量，B 中最下边的一排星号应是向量

$$(-(\alpha_k I_m-h\beta_k A)^{-1}(\alpha_0 I_m-h\beta_0 A),\cdots,\\ -(\alpha_k I_m-h\beta_k A)^{-1}(\alpha_{k-1}I_m-h\beta_{k-1}A)),$$

B 的特征值满足特征方程

$$|B-\mu I_{mk}|=0,$$

这里 I_{mk} 是 mk 阶单位矩阵. 在上边的行列式中，将第 k 列乘以 μ 加到 $k-1$ 列上，再将所得的 $k-1$ 列乘以 μ 加到 $k-2$ 列上，……，最后将所得的第二列乘以 μ 加到第一列上，这样就消掉了主对角元上边的 $k-1$ 个 $-\mu I_m$. 将最后所得的行列式展开，得

$$\left|(\alpha_k I_m-h\beta_k A)^{-1}\sum_{l=0}^{k-1}(\alpha_l I_m-h\beta_l A)\mu^l-\mu^k I_m\right|=0,$$

上式乘以 $|\alpha_k I_m-h\beta_k A|$ 则得

$$\left|\sum_{l=0}^{k}(\alpha_l I_m-h\beta_l A)\mu^l\right|=0,$$

存在非奇异矩阵 S，使 A 化为 Jordan 标准形

$$S^{-1}AS=J,$$

J 的主对角元素是 $\lambda_1, \lambda_2, \cdots, \lambda_m$，下次对角元素是 1 或零，其余元素全是零，因此以 $|S^{-1}|$ 和 $|S|$ 分别左乘和右乘上行列式，则得

$$\left|\sum_{l=0}^{k}(\alpha_l I_m-h\beta_l J)\mu^l\right|=0$$

行列式中的矩阵是下三角矩阵，故其值是主对角元素之积

$$\prod_{j=1}^{m}\left[\sum_{l=0}^{k}(\alpha_l-h\beta_l\lambda_j)\mu^l\right]$$

$$=\prod_{j=1}^{m}[\rho(\mu)-h\lambda_j\sigma(\mu)]=0,$$

共有 m 个因式，对于每个 λ_j，有一个方程

$$\rho(\mu)-h\lambda_j\sigma(\mu)=0,\ j=1,2,\cdots,m,$$

设这个方程的根是 μ_{jl} $(l=1,2,\cdots,k)$. 若所有的 mk 个根都满足

$$|\mu_{jl}|<1,\ j=1,\cdots,m,\ l=1,\cdots,k$$

则此线性多步法计算稳定.

这一结果和将一线性多步法用于试验方程

$$y'=\lambda_j y,\ j=1,\cdots,m$$

计算稳定的条件是一样的. 因而有

定理 12.2 设一阶线性常系数常微分方程组

$$y' = Ay + F(t) \tag{12.9}$$

系数矩阵 A 的特征值是 $\lambda_1, \lambda_2, \cdots, \lambda_m$. 对于步长 $h > 0$, 假若一个线性多步法用于积分试验方程 $\dot{y} = \lambda_j y$ $(j = 1, \cdots, m)$ 计算稳定,那么这个多步法用于积分上述方程组时也是计算稳定的.

我们注意, 定理 12.1 和定理 12.2 中对 (12.9) 的系数矩阵 A 没有加限制,而一般的讨论是限制 A 的特征值都是单根,即 A 的标准形中没有二维及二维以上的 Jordan 块. 这说明了试验方程有很好的代表性. 简言之, 即是一个数值积分方法若用于试验方程数值稳定,则用于方程组 (12.9) 也数值稳定.

在应用数值方法对常微分方程(组)进行积分时, 必须注意计算稳定性问题,因为破坏了计算稳定的条件,误差就会指数式的增长,淹没真解,甚至导致溢出.

有意义的自动控制系统,在线性常系数的情形, 如 (12.9), A 的特征值 λ_j 满足 $\mathrm{Re}(\lambda_j) < 0$, $j = 1, 2, \cdots, m$. 由于收敛的单步法用于试验方程 $\dot{y} = \lambda y$, 其特征值应是 $e^{h\lambda}$ 的一个近似式,所以原点邻域的左半(以虚轴为界)属于稳定区域, 而右半不属于稳定区域. 由此可知稳定区域在原点的左边, 或包括原点左边的一个区域. 收敛的线性多步法多数也有这个性质 (如 Adams 类型的内插法和外推法). 因此用这些方法去积分 (12.9) 时, 若发现计算不稳定, 只要缩小积分步长, 使 $h\lambda_j$, $j = 1, \cdots, m$, 都落到稳定区域之内时, 数值积分过程就会稳定. 通常我们叫这个做法"缩小步长的原则". 实践说明, 这个办法一般对变系数或非线性的控制系统问题, 也是适用的. 以上是指稳定区域为有界区域的情形,如显式方法就是这样. 在使用 A 稳定的方法,积分步长不再受计算稳定性的限制,而是取决于数值解的精度.

§3 右函数中避免导数的计算

在自动控制系统的框图中, 经常遇到如图 12.10 所示的环

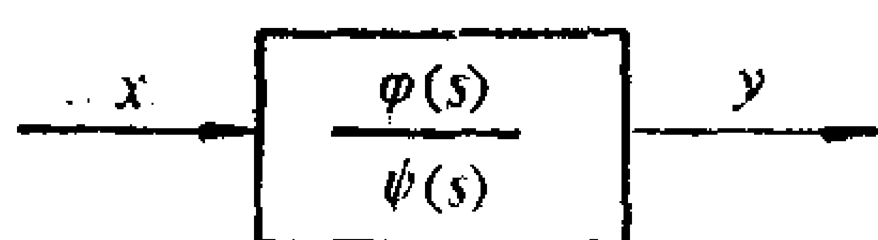


图 12.10 带有微分的环节

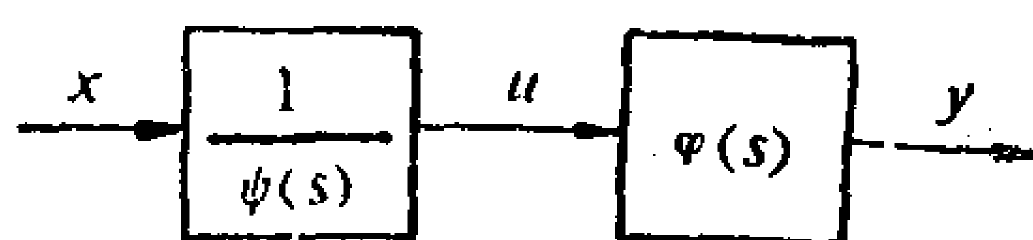


图 12.11 等价的框图

节．输入 x 和输出 y 之间的关系的微分方程形式是

$$\psi(s)y = \varphi(s)x, \tag{12.10}$$

这里 x 和 y 是 t 的函数，$s = \frac{d}{dt}$，$\varphi(s)$ 和 $\psi(s)$ 分别是 s 的 m 和 n 次实系数多项式

$$\varphi(s) = \sum_{i=0}^{m} a_i s^{m-i}, \quad \psi(s) = \sum_{i=0}^{n} b_i s^{n-i}.$$

在这一节里，除特别指明外，都假定 $\varphi(s)$ 和 $\psi(s)$ 无共因子，x 具有 m 阶连续函数．方程 (12.9) 一般是微分方程组中的一个方程，输入 x 看做已知函数时，(12.10) 对于 y 是 n 阶微分方程，要提供 n 个初值

$$y^{(k)}(0) = y_0^{(k)} \ (k = 0, 1, \cdots, n-1).$$

若 $m > 0$，在数值积分 (12.10) 时，需要计算 x 的直到 m 阶的导数，用数值方法求导数是很困难的，因为求微分的步长大了，会产生较大的截断误差，步长小了，又会严重损失有效数字，一般说，一次数值求导，就要损失近一半的有效数字．更严重的是，在右函数中做数值求导时，常常会引起数值积分过程计算不稳定，使计算无法进行．因此应该找到一种算法，避免求 x 的导数．

看来这也是可能的．概括地说，在解微分方程 (12.10) 时，先要对 x 微分，然后又要积分得到 y，能否将这两种相逆的运算"抵消"呢？在 [15] 中将框图 12.10 改成等价的框图 12.11，引进中间变量 u，相应的方程是

$$\psi(s)u = x, \tag{12.11}$$

$$y = \varphi(s)u. \tag{12.12}$$

在 $m \leqslant n$ 时，由 (12.10) 可以解出 u，在解 u 的过程中就算出了 $u^{(k)}$，$k = 1, \cdots, n$，将这些值代入 (12.11) 就得出了 y 在该节点

的值.

解 (12.10) 需要 u 的初值 $u(0),\cdots,u^{(n-1)}(0)$, 可以如下求法. 对方程 (12.10) 微分 k 次, $k=0,1,\cdots,m-1$, 对方程 (12.11) 微分 l 次, $l=0,1,\cdots,n-1$, 将 $t=0$ 代入各式, 则得到 $u(0),\cdots,u^{(m+n-1)}(0)$ 的 $m+n$ 个方程, 其右端项是 $x^{(k)}(0)$, $k=0,1,\cdots,m-1$, $y^{(l)}(0)$, $l=0,1,\cdots,n-1$, 都是已知的值, 解这个线性代数方程组就可以得到 $u(0),\cdots,u^{(n-1)}(0)$. 只要 $\varphi(s)$ 和 $\psi(s)$ 没有公因子, 方程组的系数行列式就不等于零, 方程组有唯一解.

若 $m\geqslant n$, 以 $\psi(s)$ 除 $\varphi(s)$

$$\varphi(s)=Q(s)\psi(s)+R(s),$$

其中 $Q(s)$ 的次数是 $m-n$, $R(s)$ 的次数是 r, $r\leqslant n-1$. 将它代入 (12.12), 并用 (12.11), 得

$$y=Q(s)x+R(s)u, \tag{12.13}$$

其框图如图 12.12 所示.

计算 u 的初值时, 将 (12.11) 微分 k 次, $k=0,1,\cdots,r-1$, 将 (12.12) 微分 l 次, $l=0,1,\cdots,n-1$, 将 $t=0$ 代入各方程, 就得 $n+r$ 个未知数 $u(0),\cdots,u^{(n+r-1)}(0)$ 的 $n+r$ 个方程, 解这个线性代数方程组得 $u(0),\cdots,u^{(n-1)}(0)$. 方程组中用到 $y(0),\cdots,y^{(n-1)}(0)$ 和 $x(0),\cdots,x^{(m-1)}(0)$.

总之, 若 $m\leqslant n$, 将积分 (12.10) 换成积分 (12.11), 并计算 (12.12). 在积分过程中可以完全避免导数计算, 但在求 u 的初值时要用到 $x(0),\cdots,x^{(m-1)}(0)$. 若 $m>n$、将积分 (12.10) 换成积分 (12.11), 并计算 (12.13), 这时要用到 $x(t),\cdots,x^{(m-n)}(t)$, 初值计算中要用到 $x(0),\cdots,x^{(r-1)}(0)$, $r\leqslant n-1$.

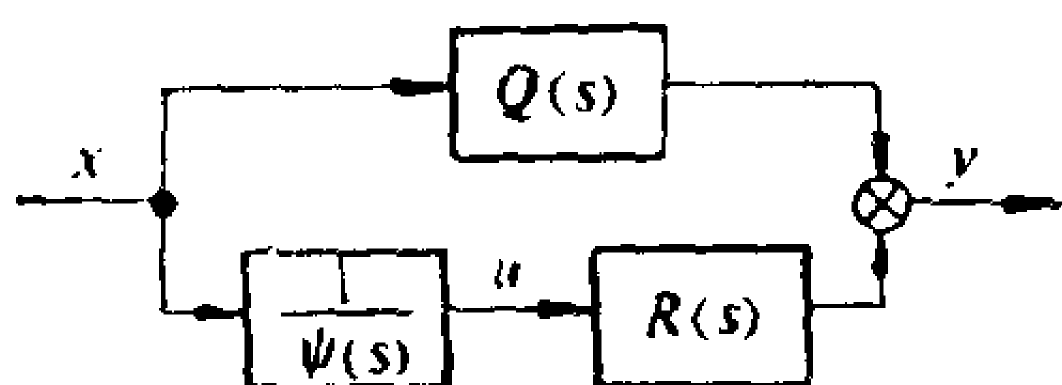


图 12.12 $m\geqslant n$ 的等价框图

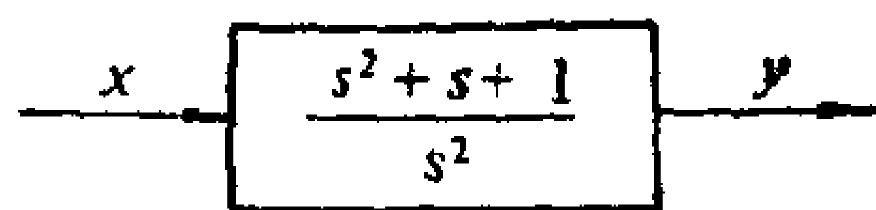


图 12.13 一个带有微分的环节

应当指出，自动控制所用的元件，如放大器，传感器，阻尼器，执行部件等，多数是一阶或二阶的，即 $n \leqslant 2$，并且多数是 $m < n$，即积分过程和 u 的初值计算都不出现求导问题. 在 $m = n = 2$ 的情形，若将 (12.10) 换成 (12.11) 和 (12.13)，这时 $Q(s)$ 是一个常数，计算 u 的初值时要用 $\dot{x}(0)$.

在 [15] 中还指出，变换之后，对 x 有第一类间断的情形(即在间断点左导数和右导数都存在)也能适用. 例如有图 12.13 中的环节，这里

$$x = \begin{cases} 0 & \text{当 } t < 1, \\ 1 & \text{当 } t \geqslant 1, \end{cases}$$

$$y_0 = y_0' = 0,$$

其方程为

$$y'' = x'' + x' + x,$$

方程的真解是

$$y = \begin{cases} 0 & \text{当 } t < 1, \\ t + (t-1) + \frac{1}{2}(t-1)^2 & \text{当 } t \geqslant 1, \end{cases}$$

将方程变为 (12.11), (12.13) 的形式，其框图如 12.14 所示.

$$u'' = x,$$

$$y = x + u' + u,$$

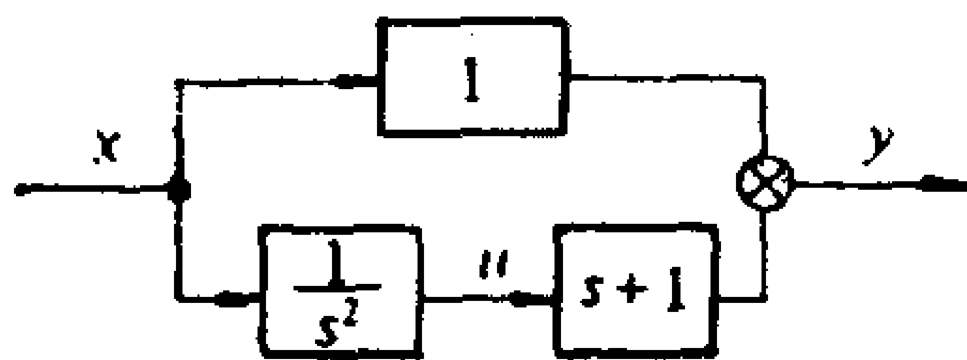


图 12.14 等价框图

容易算出 u 的初值

$$u(0) = u'(0) = 0.$$

用数值方法积分上述第一方程时，得到如下的 u 的一个近似

$$u = \begin{cases} 0 & \text{当} t < 1, \\ \frac{1}{2}(t-1)^2 & \text{当 } t \geqslant 1, \end{cases}$$

代入第二个方程则得到真解的一个近似.

而直接解原来的方程,在 $t=1$ 时 x'', x' 的值无法表示,若取的步长避开这个点,实际上等于略去了 x'' 和 x', 所得的数值解是如下的函数 z 的近似

$$z=\begin{cases}0 & \text{当 } t<1\\ \frac{1}{2}(t-1)^2 & \text{当 } t\geqslant 1\end{cases}$$

这是根本性的错误. 以上的例子是最简单的情形, 目的是揭示问题的实质.

下边我们介绍解决这个问题的另一种处理方法*. 以二阶微分方程

$$y''+a_1y'+a_2y=b_0x''+b_1x'+b_2x+f(t) \qquad (12.14)$$

为例介绍其处理的思想. 令

$$u=y'+a_1y-b_0x'-b_1x, \qquad (12.15)$$

函数 u 中的各项,是 (12.14) 中可以积分出来的项. 对 (12.15) 求导,代入 (12.14), 得

$$u'=-a_2y+b_2x+f(t), \qquad (12.16)$$

再令 (12.15) 中可以积分出来项为 v, 即

$$v=y-b_0x, \qquad (12.17)$$

将其求导,代入 (12.15), 得

$$v'=u-a_1y+b_1x. \qquad (12.18)$$

(12.16), (12.18) 就是我们要求的微分方程组, 其中的 y 可以用 (12.17) 消去. 解出 u, v 之后,可通过 (12.17) 求 y.

u, v 的初值可将 $t=0$ 代入(12.15),(12.17) 得到

$$u_0=y_0'+a_1y_0-b_0x_0'-b_1x_0,$$
$$v_0=y_0-b_0x_0.$$

在 [8] 中提出了第三种处理方法. 将形状为 (12.10) 的方程表示成

* 这个方法是高永春 1965 年提出的.

$$(a_0 + a_1 s + \cdots + a_{n-1}s^{n-1} + a_n s^n)y$$
$$= (b_0 + b_1 s + \cdots + b_m s^m)x \tag{12.19}$$

首先考虑 $m = n$ 的情形．将 (12.19) 写成

$$s^n(y - b_n x) + s^{n-1}(a_{n-1}y - b_{n-1}x) + \cdots$$
$$+ s(a_1 y - b_1 x) + a_0 y - b_0 x = 0,$$

令

$$y_1 = y - b_n x, \tag{12.20}$$

$$y_{j+1} = y'_j + a_{n-j}y - b_{n-j}x, \quad j = 1, \cdots, n-1. \tag{12.21}$$

原方程化为

$$y'_n + a_0 y - b_0 x = 0, \tag{12.22}$$

将 (12.21) 和 (12.22) 合在一起，得到与 (12.19) 等价的方程组

$$y'_j = -a_{n-j}y_1 + y_{j+1} + (b_{n-j} - a_{n-j}b_n)x,$$
$$j = 1, \cdots, n, \tag{12.23}$$

其中 $y_{n+1} = 0$．由 (12.20)，原方程 (12.19) 的解为

$$y = y_1 + b_n x.$$

方程组 (12.23) 可写成矩阵形式

$$\begin{pmatrix} y'_1 \\ y'_2 \\ \vdots \\ y'_{n-1} \\ y'_n \end{pmatrix} = \begin{pmatrix} -a_{n-1} & 1 & & & \\ -a_{n-2} & 0 & 1 & & \\ \vdots & & & \ddots & \\ -a_1 & & & & 1 \\ -a_0 & & & & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}$$

$$+ x \begin{pmatrix} b_{n-1} - a_{n-1}b_n \\ b_{n-2} - a_{n-2}b_n \\ \vdots \\ b_1 - a_1 b_n \\ b_0 - a_0 b_n \end{pmatrix},$$

其初值为

$$\begin{pmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \\ \vdots \\ y_n(0) \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ a_{n-1} & 1 & & & \\ a_{n-2} & a_{n-1} & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ a_1 & \cdots & a_{n-2} & a_{n-1} & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ y'_0 \\ y''_0 \\ \vdots \\ y^{(n-1)} \end{pmatrix}$$

$$
-\begin{pmatrix} b_n & & & \\ b_{n-1} & b_n & & \\ b_{n-2} & b_{n-1} & b_n & \\ \vdots & \ddots & \ddots & \ddots \\ b_1 & \cdots & b_{n-2} & b_{n-1} & b_n \end{pmatrix} \begin{pmatrix} x_0 \\ x_0' \\ x_0'' \\ \vdots \\ x_0^{(n-1)} \end{pmatrix}. \tag{12.24}
$$

对于$m < n$的情形，只要在以上的讨论中令 $b_{m+1}, \cdots, b_n = 0$，就可得到所需的结果.

对于$m > n$的情形，令

$$
x^{(m-n)} = \bar{x},\ b_{m-n+k} = \bar{b}_k\ (k = 0, 1, \cdots, m),
$$

(12.19) 变为

$$
s^n(y - \bar{b}_n\bar{x}) + s^{n-1}(a_{n-1}y - \bar{b}_{n-1}\bar{x}) + \cdots + s(a_1 y - \bar{b}_1\bar{x})
$$
$$
+ a_0 y - \bar{b}_0\bar{x} + (b_{m-n-1}s^{(m-n-1)} + \cdots + b_0)x,
$$

作同样的变换，相当于 (12.22) 的方程是

$$
y_n' = -a_0 y + \bar{b}_0\bar{x} + (b_{m-n+1}s^{m+n+1} + \cdots + b_0)x,
$$

所得的等价方程组是

$$
\begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_{n-1}' \\ y_n \end{pmatrix} = \begin{pmatrix} -a_{n-1} & 1 & & 0 \\ -a_{n-2} & 0 & 1 & \\ \vdots & & & \ddots \\ a_1 & & 0 & & 1 \\ a_0 & & & & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}
$$

$$
+ x^{(m-n)} \begin{pmatrix} b_{m-1} - a_{n-1}\ b_m \\ b_{m-2} - a_{n-2}\ b_m \\ \vdots \\ b_{m-n+1} - a_1\ b_m \\ b_{m-n} - a_0\ b_m \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \sum_{j=0}^{m-n+1} b_j\, x^{(j)} \end{pmatrix}.
$$

原方程的解是

$$
y = y_1 + b_m x^{(m-n)},
$$

方程组的初值是

$$
\begin{pmatrix} y_1(0) \\ y_2(0) \\ y_3(0) \\ \vdots \\ y_n(0) \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ a_{n-1} & 1 & & 0 & \\ a_{n-2} & a_{n-1} & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ a_1 & \cdots & a_{n-2} & a_{n-1} & 1 \end{pmatrix} \begin{pmatrix} y_0 \\ y_0' \\ y_0'' \\ \vdots \\ y^{(n-1)} \end{pmatrix}
- \begin{pmatrix} b_m & & & & \\ b_{m-1} & b_m & & 0 & \\ b_{m-2} & b_{m-1} & b_m & & \\ \vdots & \ddots & \ddots & \ddots & \\ b_{m-n+1} & \cdots & b_{m-2} & b_{m-1} & b_m \end{pmatrix} \begin{pmatrix} x_0^{(m-n)} \\ x_0^{(m-n+1)} \\ x_0^{(m-n+2)} \\ \vdots \\ x_0^{(m-1)} \end{pmatrix}.
$$

后两种处理方法，对 x 有第一类间断点都能适用．最后一种处理方法比较规则，方程组的初值计算容易．

§4 框图的变换

如上节第一种处理方法，若一种框图对应的微分方程组数值解有困难，有时换成等价的框图，所对应的方程组求数值解可能就容易一些，这种等价变换可以自动控制的观点为依据，也可以数学的观点为依据．以下我们举出一个从数学观点来变换框图的例子[17]．

在自动控制迴路中有时出现如图 12.15 所示的带有微分反馈的限幅环节．这里 x 是输入，y 是输出，v 是反馈，z 是中间变量，

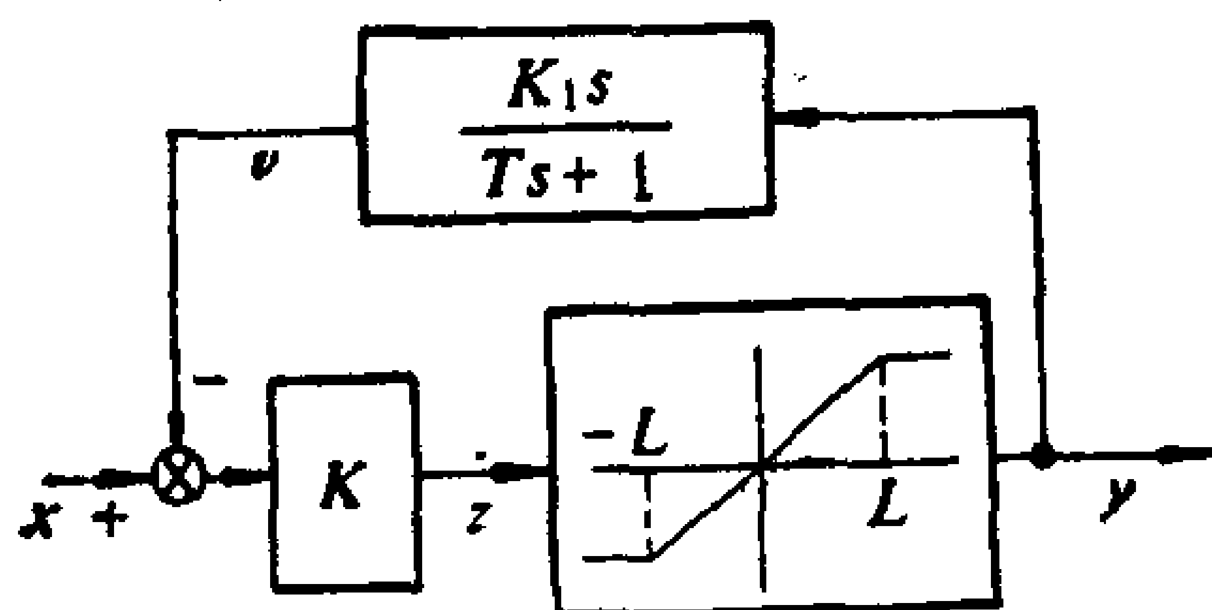


图 12.15　带有微分反馈的限幅环节

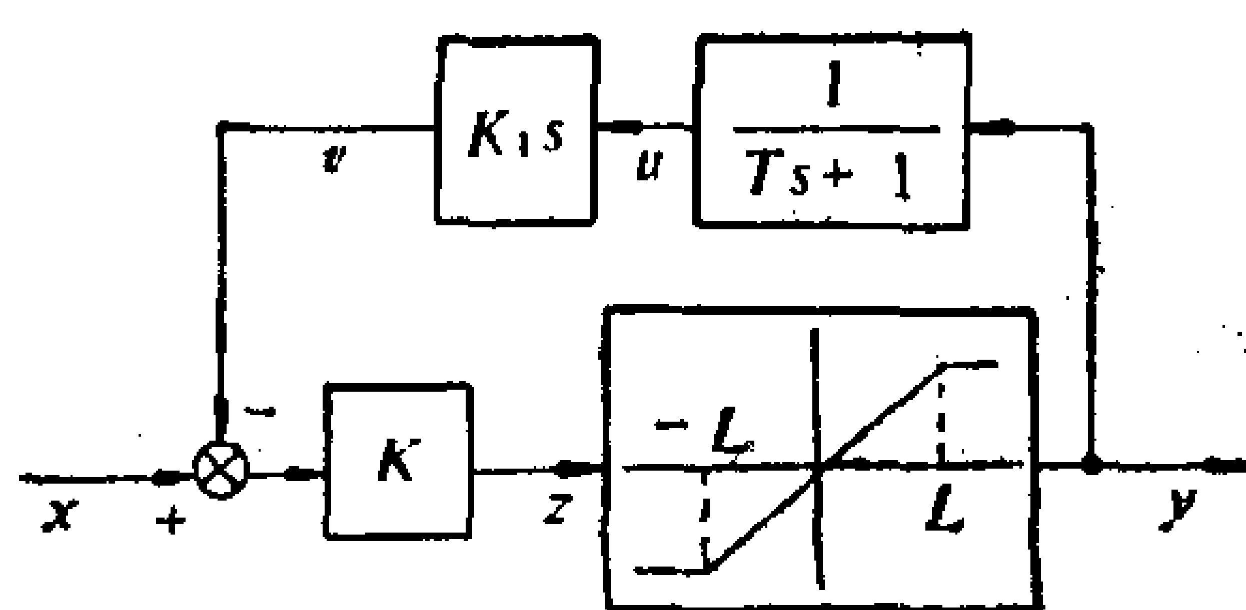


图 12.16　初步处理后的框图

$s=\frac{d}{dt}$, K, K_1 是放大倍数，T 是时间常数，K, K_1, T 都是正的常数．

反馈回路对应的微分方程中右端函数含有导数 y'，我们用上节的第一种方法处理，则得如图 12.16 所示的框图，相应的方程组是

$$Tu'+u=y, \tag{12.25}$$

$$v=K_1u', \tag{12.26}$$

$$z=K(x-v), \tag{12.27}$$

$$y=\begin{cases} z, & |z|\leqslant L, \\ L\operatorname{sign}z, & |z|>L. \end{cases} \tag{12.28}$$

这里 u 是新引进的中间变量．方程组中 x 是输入，看做已知函数．原来含有 v 的微分方程，应当提供初值 v_0，解上述方程组时应算出初值 u_0． 可以如下计算：已知 v_0 由 (12.26) 算出 u_0'，由 (12.27) 算出 z_0，将 z_0 代入 (12.28) 算出 y_0，将 u_0'，y_0 代入 (12.25) 算出 u_0，其计算步骤如图 12.17 所示．

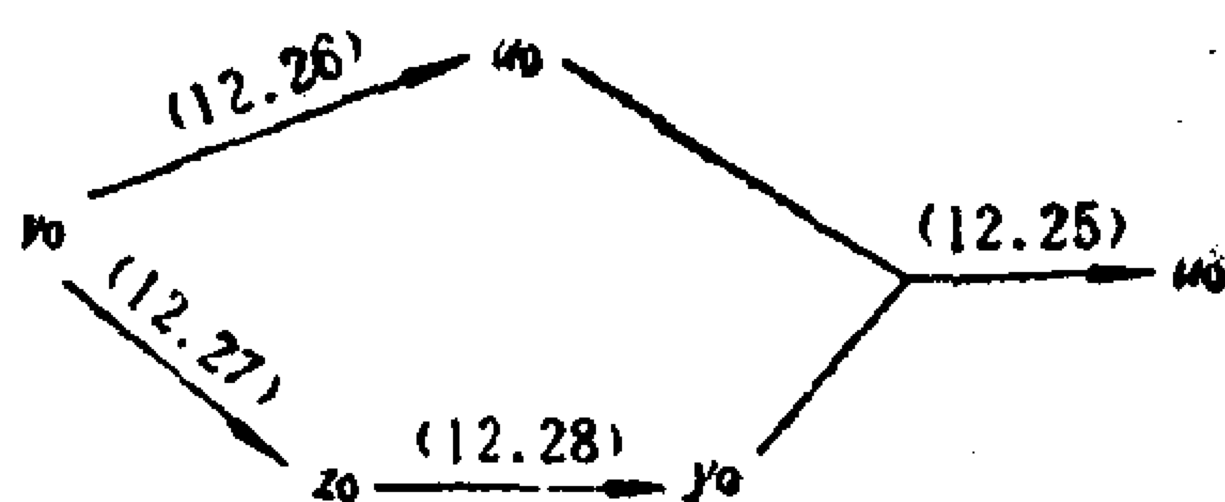


图 12.17 u_0 的计算次序

在数值积分时，应该从 $t=t_n$ 时 x, u 的值 x_n, u_n 算出 u_n' 的值．但是由 (12.25) 算 u_n' 时需要知道 y_n，由 (12.28) 算 y_n 时需要知道 z_n，由 (12.27) 算 z_n 时需要知道 v_n，从 (12.26) 算 v_n 时又需要知道 u_n'，其循环关系如图 12.18 所示．也就是说，应该将四个方程看做一个代数方程组，将 u_n' 解出来．由于 (12.28) 是非线性关系，解方程组也有困难．

引进等价的框图，其想法是，假定相应于 (12.28) 的限幅器不

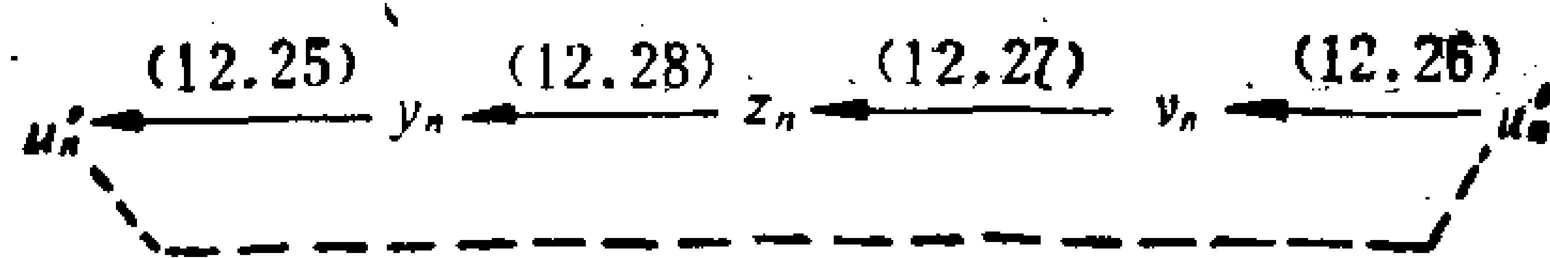


图 12.18 求 u'_n 的循环关系

限幅,即 $y = z$, 再和(12.25),(12.26),(12.27)一起消去 u', v, z 后解出 y, 因为它不是真正的 y, 将它表做 η, 将 η 限幅后得到 y. 于是计算的公式变成

$$Tu' + u = y, \tag{12.25}$$

$$\eta = \frac{KT}{T + KK_1}\left(x + \frac{K_1}{T}u\right), \tag{12.29}$$

$$y = \begin{cases} \eta, & 若\ |\eta| < L, \\ L\mathrm{sign}\eta, & 若\ |\eta| > L, \end{cases} \tag{12.30}$$

其框图如图 12.19 所示.

在[17]中证明了两个框图的等价性,解的存在性和唯一性.在开始时,从 v_0 算 u_0, 用图 12.17 对应的过程计算. 在积分中,知道 u_n 算 u'_n, 用图 12.19 所指的过程计算, 即从(12.29)算 η_n, 从(12.30)算 y_n, 最后由(12.25)即可算出 u'_n.

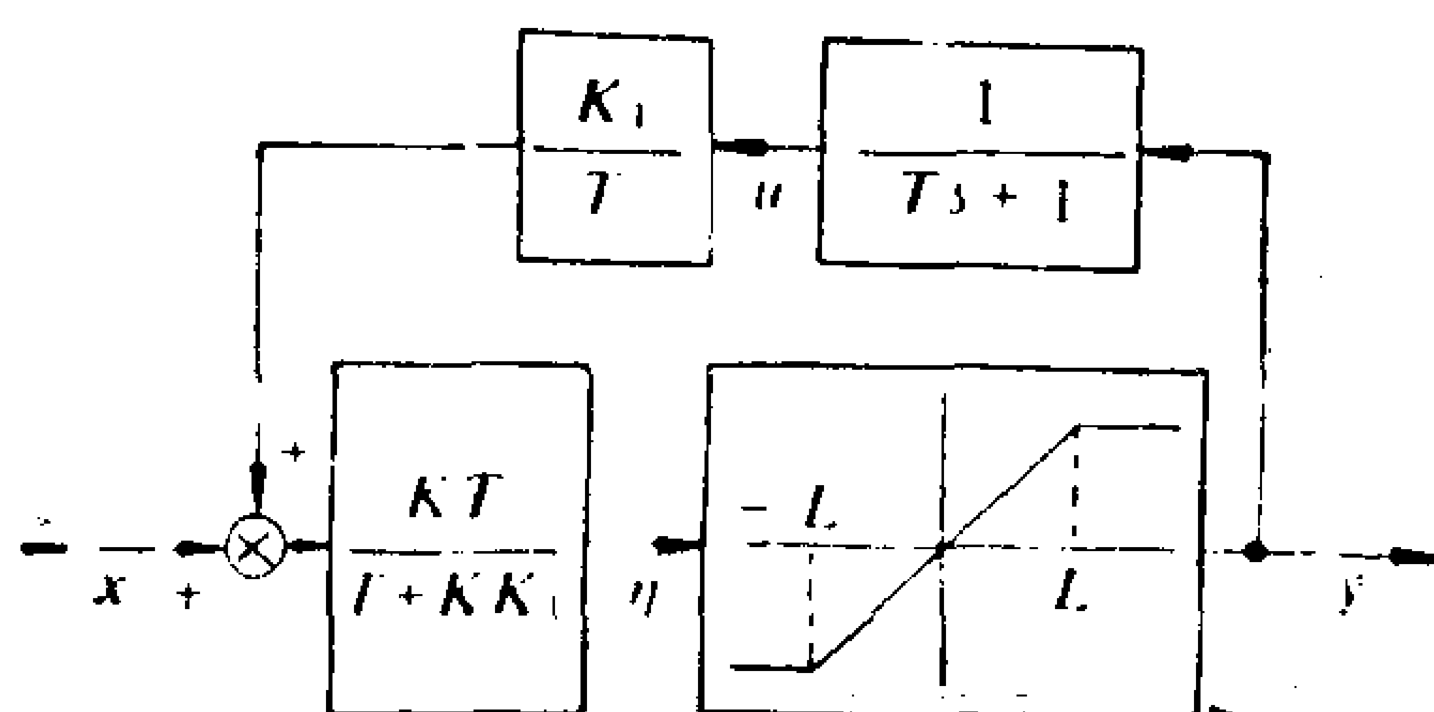


图 12.19 变换后的框图

§5 非正规格式的计算稳定性

由给出的 t 和 y 的值,计算相应的 y'的值,代入数值公式中去积分,我们把这种计算格式叫做正规格式,否则叫非正规格式. 换言之,正规格式是严格按数值积分公式要求计算的格式,没有按数

值积分公式要求的计算格式是非正规格式。非正规格式有不同的计算稳定性要求，应具体分析，为此我们举两个简化了的例子。

例 1　设微分方程为

$$y' = f(t, y, z),$$

而 z, y 满足方程

$$z = g(y, z).$$

设用 Euler 公式，在 t_n 点知道 y_n 后，本来可以从第二个方程解出 z_n 来，然后由第一个方程计算 y'_n，进行积分。若第二个方程非常复杂，解 z_n 有困难，有时用 y_n 和 z_{n-1} 代入 g 中求 z，取作 z_n。这就是一种非正规格式

$$y_{n+1} = y_n + hf(t_n, y_n, z_n),$$
$$z_{n+1} = g(y_{n+1}, z_n).$$

误差的线性化方程是

$$\begin{pmatrix} 1 & 0 \\ -g_y & 1 \end{pmatrix} \begin{pmatrix} \triangle y_{n+1} \\ \triangle z_{n+1} \end{pmatrix} = \begin{pmatrix} 1 + hf_y & hf_z \\ 0 & g_z \end{pmatrix} \begin{pmatrix} \triangle y_n \\ \triangle z_n \end{pmatrix},$$

迭代矩阵的特征方程是

$$\lambda^2 - (1 + g_z + hf_y + hf_z g_y)\lambda + (1 + hf_y)g_z = 0,$$

当 $h \to 0$ 时，两个特征根的极限分别是 1 和 g_z。所以这个格式计算稳定的必要条件是 $|g_z| < 1$。

例 2　设微分方程为

$$y' = f(t, y, y'),$$

以这个方程做为右函数含有导数的例子。在 t_n 知 y_n 后本应解出 y'_n，然后积分。若方程过于复杂，解 y'_n 有困难，用数值微分 $(y_n - y_{n-1})/h$ 代替函数 f 中的 y'_n，设以步长为 h 的 Euler 法解之

$$y_{n+1} = y_n + hf(t_n, y_n, (y_n - y_{n-1})/h),$$

误差满足的线性化方程为

$$\triangle y_{n+1} = \triangle y_n + hf_y \triangle y_n + f_{y'}(\triangle y_n - \triangle y_{n-1}).$$

将 f_y, $f_{y'}$ 看成常数，其特征方程为

$$\lambda^2 - (1 + hf_y + f_{y'})\lambda + f_{y'} = 0,$$

当 $h \to 0$ 时，两根的极限分别是 1 和 $f_{y'}$. 所以计算稳定的必要条件是 $|f_{y'}| < 1$.

我们指出，相应于图 12.15 的反馈迴路的方程是

$$Tv' + v = K_1 y',$$

若 y'_n 用数值微分 $(y_n - y_{n-1})/h$ 代替，假定限幅只工作在线性区域，即 $y = z$，若用步长为 h 的 Euler 法解此方程组，则对齐次方程组有

$$T\frac{v_{n+1} - v_n}{h} + v_n = K_1 \frac{y_n - y_{n-1}}{h},$$

$$z_n = K(-v_n), \quad y_n = z_n,$$

消去 y 和 z，得

$$v_{n+1} + \left(\frac{KK_1}{T} - 1 + \frac{h}{T}\right) v_n - \frac{KK_1}{T} v_{n-1} = 0,$$

当 $h \to 0$，其特征根是 1 和 $-\frac{KK_1}{T}$. 故当 $KK_1 > T$ 时计算不稳定，K 和 K_1 是放大系数，T 是时间常数，一般都有 $KK_1 > T$，即一般是计算不稳定的.

§6 其他问题的处理

在自动控制系统的运动方程中，有时含有不是微分方程形式的方程，最好是将它化为微分方程形式，可以并入微分方程组一起做数值积分. 例如

1. 形式为

$$y(t) = y_0 - \int_0^t f(\tau) d\tau$$

的方程，可以对 t 微分，化为微分方程

$$\dot{y} = -f(t), \quad y(0) = y_0.$$

2. 形式为

$$y(t) = \int_0^t K(\tau) \cos(t - \tau) d\tau$$

的方程,将 $\cos(t-\tau)$ 展开

$$\cos(t-\tau)=\cos t\cos\tau+\sin t\sin\tau,$$

代入方程

$$y(t)=\cos t\int_0^t K(\tau)\cos\tau d\tau+$$
$$\sin t\int_0^t K(\tau)\sin\tau d\tau.$$

令

$$z_1'=K(t)\cos t,\ \ z_1(0)=0,$$
$$z_2'=K(t)\sin t,\ \ z_2(0)=0,$$

则有

$$y(t)=z_1(t)\cos t+z_2(t)\sin t,$$

z_1 和 z_2 的两个微分方程可以加入微分方程组积分.

3. 如在右端函数中需要解方程

$$F(t,x,y)=0,$$

这里 t 是自变量，x 是积分中提供的值，例如是一个微分方程的解,这时 x' 也是可以算出来的,y 是要解方程求出的值. 我们可以将 x 和 y 都看做 t 的函数[16,17]，对方程求导则得

$$F_t+F_xx'+F_yy'=0.$$

若 x' 是可以算出来的,则可在方程组中加一个方程

$$y'=-(F_t+F_xx')/F_y.$$

在解微分方程的过程中就可求出 y 来. 当然要假定在解微分方程的过程中 $F_y\neq 0$，这个要求是自然的. 还要给出 y 的初值,这是解微分方程所要求的.

如果所解的微分方程中,有一个函数指数衰减,可以选变化的比例因子,例如

4. 有方程

$$y'+\lambda y=f(t,y),$$

这里 $\mathrm{Re}(\lambda)<0$，$f(t,y)$ 是很小的量，可取 $e^{\lambda t}$ 做为比例因子. 以 $e^{\lambda t}$ 乘方程两边

$$\frac{d}{dt}[e^{\lambda t}y] = e^{\lambda t}f(t, y),$$

令 $z = e^{\lambda t}y$，则有

$$z' = e^{\lambda t}f(t, ze^{-\lambda t}).$$

以下的两种处理，是方程中的近似，是否可用，要从物理上考虑，或以计算结果对比，判断其是否可用。

5. 略去小参数。例如对放大器环节的方程

$$Ty' + y = Kx,$$

当时间常数 T 很小时，有时可以略去。这就相当于忽略了暂态过程。这是自动控制系统计算中常用的一种方法。若有可能，有时可使求解难度减小。

6. 时延处理。如有时延环节

$$y(t) = x(t - \tau),$$

这里 τ 是小的正常数。在方程组数值积分时，可以保留 x 的值，进行插值，也可以变成

$$y(t + \tau) = x(t),$$

将 $y(t + \tau)$ 展开，略去 τ 的高阶项，得出一个微分方程，参加方程组积分。如略去 τ 的二阶项，得

$$\tau y' + y = x(t), \tag{12.31}$$

就变成了一个放大环节。如果略去 τ 的三阶项，得

$$\frac{1}{2}\tau^2 y'' + \tau y' + y = x(t) \tag{12.32}$$

变成了振荡环节。若 $x(t)$ 是阶跃函数，$y(t) \equiv 0, t < 0$，方程 (12.31) 的解是

$$y = 1 - e^{-\frac{1}{\tau}t}.$$

而方程 (12.32) 的解是

$$y = 1 - \sqrt{2}\, e^{-\frac{1}{\tau}t}\sin\left(\frac{1}{\tau}t + \frac{\pi}{4}\right),$$

函数 x，真解 y，方程 (12.31) 和方程 (12.32) 的解如图 12.20—12.23 所示。

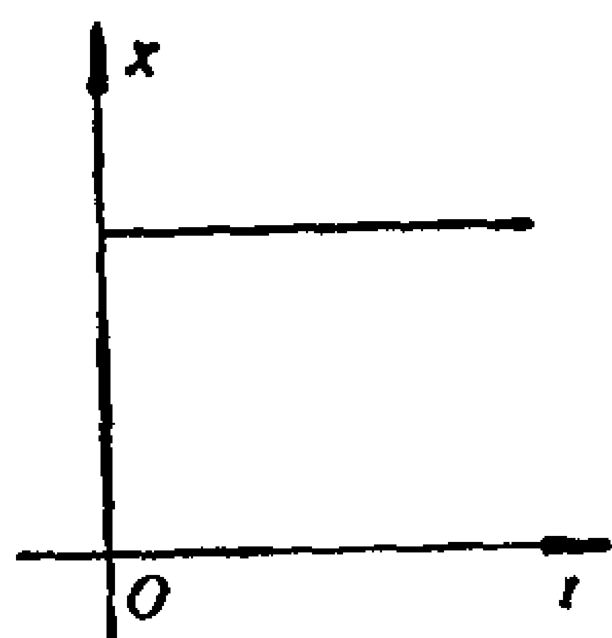


图 12.20　x 的图形

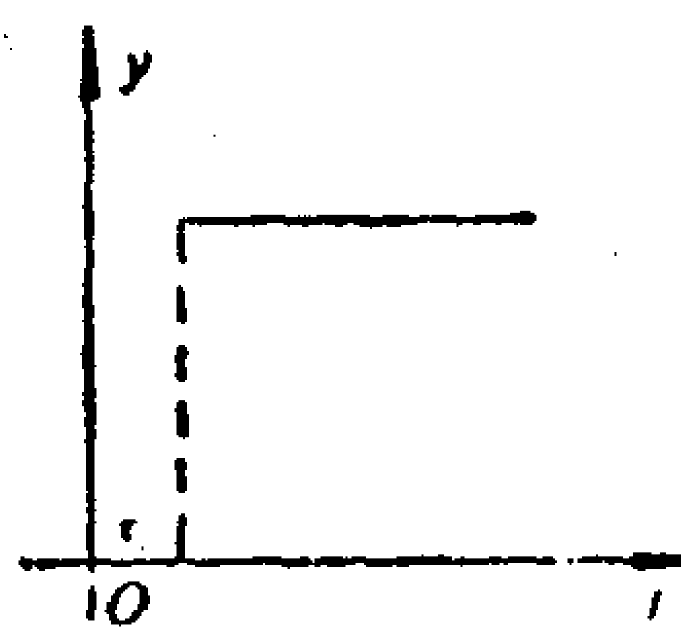


图 12.21　真解的图形

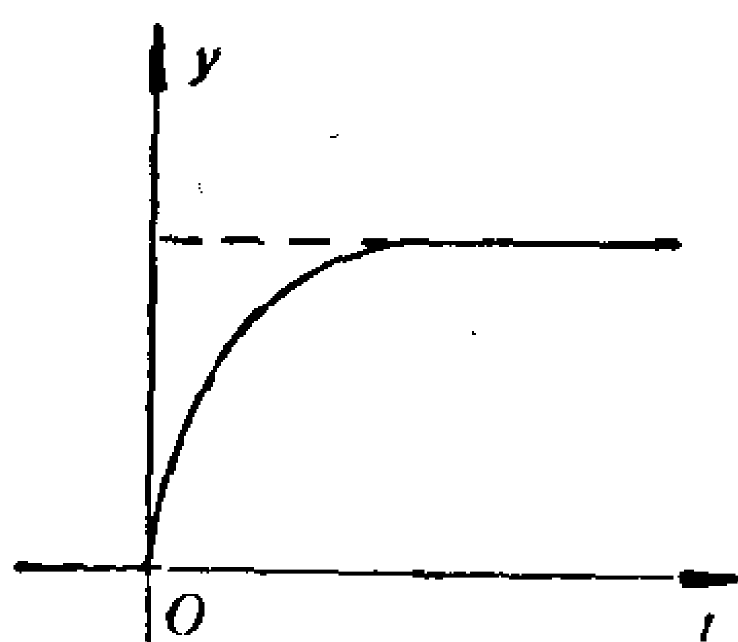


图 12.22　方程(12.31)的解

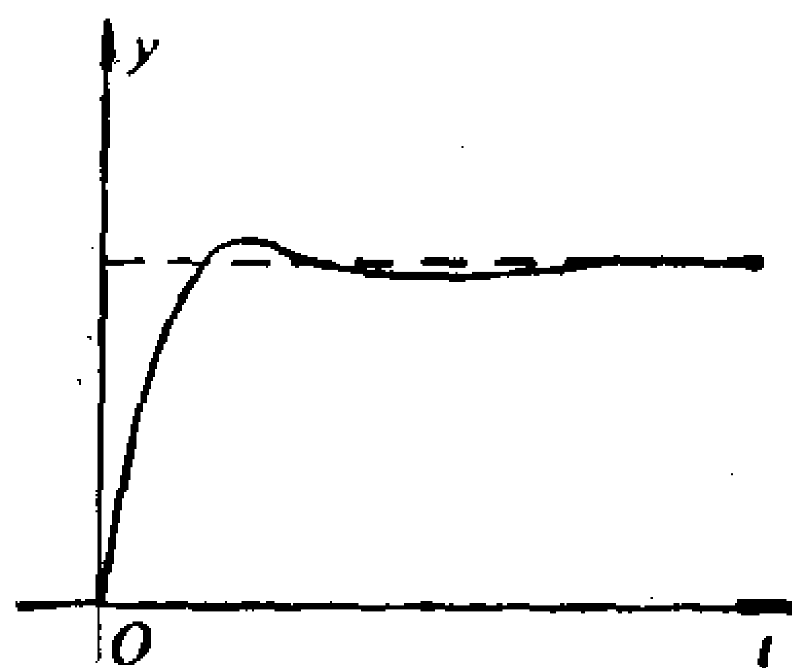


图 12.23　方程(12.32)的解

第十三章　处理刚性方程的一些其它方法

这一章简短地叙述一些处理刚性方程的方法，主要是提供一些处理的思想.

§1　等效系统替代方法

用传统的数值方法求解刚性方程产生困难的一个基本原因是不能很好地拟合方程的解，而且这些拟合误差的传播是不稳定的，误差被方法放大了. 如果我们能直接得到方程的解析解，再由解析解求得具体数值，这些问题就不存在了. 但是对于一般的系统寻找解析解是困难的. 许多人提出利用等效系统代替原来系统的思想. 具体地说，假使我们要求刚性方程

$$y' = f(t, y),\ \ y(0) = y_0 \tag{13.1}$$

的解，可以不直接考虑 (13.1)，而来求 (13.1) 的等效系统

$$y' = g(t, y, u(t)),\ \ y(0) = y_0 \tag{13.2}$$

的解，其中 $u(t)$ 为可控参数. 所谓 (13.2) 是 (13.1) 的等效系统是指可以找到控制参数 $u(t)$，使得 (13.2) 的解与 (13.1) 的解一致. 如果 (13.2) 是非刚性的，或者虽是刚性的，但可用解析解来处理的话，我们就可以由 (13.1) 转换成考虑 (13.2)，从而避开了刚性引起的数值积分的困难.

韩天敏[5] 应用了类似的思想，推导了一类只须应用简单迭代的计算格式，从而可以避免计算右函数的 Jacobi 矩阵和矩阵求逆等复杂的运算. 这对于求解高阶方程组是特别有益的. 这一节我们介绍韩天敏构造的方法.

考虑初值问题

$$\begin{cases}E(t, y)y' = f(t, y), \\ y(0) = y_0,\end{cases} \tag{13.3}$$

其中 $E(t, y)$ 是m阶矩阵，y，y_0 和 f 均是m维向量．引入等效线性系统

$$By' + y = u(t), \tag{13.4}$$

其中B是m阶对角型矩阵，对角线上是任意指定的正常数．$u(t)$ 即为 (13.2) 中的控制变量．要求选取 $u(t)$ 使 (13.4) 和 (13.3) 具有相同的物理效果．如果经过离散化处理，在 $t = h, 2h, \cdots, nh, \cdots$ 的节点上能求出 $u(h)$，$u(2h), \cdots, u(nh), \cdots$的值，我们便可以从 (13.4) 出发去构造积分格式，从而克服数值求解时的困难．

将 (13.3) 和 (13.4) 式相减得

$$(E - B)y' = g - u, \tag{13.5}$$

其中 $g = f(t, y) + y$，$E = E(t, y)$，$u = u(t)$，从而使 (13.3) 和 (13.4) 式建立了联系．令

$$z = (E - B)^{-1}u,$$

则 (13.4) 和 (13.5) 式分别变为

$$y' = -B^{-1}y + (B^{-1}E - I)z, \tag{13.6}$$

$$y' = (E - B)^{-1}g - z, \tag{13.7}$$

其中I为单位矩阵．方程 (13.6)、(13.7) 中包含两个未知数，使我们可能在离散节点上用近似办法把 $z(h)$，$z(2h), \cdots, z(nh), \cdots$ 求出来．由于 $z(t) = (E - B)^{-1}u(t)$，亦即把 $u(h)$，$u(2h), \cdots, u(nh), \cdots$求出来．为此，消去 (13.6)、(13.7) 中的 y'，并用 $t = t_{n+1}$ 代入，$t = t_{n+1}$ 的函数值以下标表示，得恒等式

$$y_{n+1} + B(E_{n+1} - B)^{-1}g_{n+1} = E_{n+1}z_{n+1} \tag{13.8}$$

对 (13.7) 的右端第一项用矩形公式，第二项用梯形公式近似积分得

$$y_{n+1} = y_n + h[E_n - B]^{-1}g_n - \frac{h}{2}z_n - \frac{h}{2}z_{n+1} + O(h^2) \tag{13.9}$$

令 $x_n = \frac{h}{2} z_n$，并略去 (13.9) 式中的 $O(h^2)$ 项，(13.8) 和 (13.9) 式变成

$$y_{n+1} + B(E_{n+1} - B)^{-1} g_{n+1} = \frac{2}{h} E_{n+1} x_{n+1},$$

$$y_{n+1} = y_n + h[E_n - B]^{-1} g_n - x_n - x_{n+1}.$$

由此求得

$$x_{n+1} = \left(I + \frac{2}{h} E_{n+1}\right)^{-1} \{B(E_{n+1} - B)^{-1} g_{n+1} + y_n + h[E_n - B]^{-1} g_n - x_n\}, \tag{13.10}$$

有了 x_{n+1} 之后，就可以用 (13.4) 式或等价地用 (13.6) 式去构造积分格式.

(13.6) 式的真解的解析形式为

$$y(t) = e^{-B^{-1}(t-t_n)} y_n + \int_{t_n}^{t} e^{-B^{-1}(t-\xi)} (B^{-1} E(\xi, y(\xi)) - I) z(\xi) d\xi, \tag{13.11}$$

对 (13.11) 式中的积分使用梯形公式，则

$$y_{n+1} = e^{-B^{-1}h} y_n + e^{-B^{-1}h} (B^{-1} E_n - I) x_n + (B^{-1} E_{n+1} - I) x_{n+1}. \tag{13.12}$$

(13.10) 和 (13.12) 组成确定 x_{n+1} 和 y_{n+1} 的联立方程. 在用迭代法求解这个联立方程时，为了获得较好的初值，可以对 (13.11) 中的积分采用矩形公式，得 y_{n+1} 的预估式

$$y_{n+1}^{(0)} = e^{-B^{-1}h} y_n + 2e^{-B^{-1}h} (B^{-1} E_n - I) x_n,$$

从而建立了如下的迭代格式 (I)

$$\mathrm{I} \begin{cases} \text{(i)} \; y_{n+1}^{(0)} = e^{-B^{-1}h} y_n + 2e^{-B^{-1}h} (B^{-1} E_n - I) x_n, \\ \text{(ii)} \; x_{n+1}^{(l+1)} = \left(I + \frac{2}{h} E_{n+1}^{(l)}\right)^{-1} \{B(E_{n+1}^{(l)} - B)^{-1} g_{n+1}^{(l)} \\ \qquad + y_n + h(E_n - B)^{-1} g_n - x_n\}, \\ \text{(iii)} \; y_{n+1}^{(l+1)} = e^{-B^{-1}h} y_n + e^{-B^{-1}h} (B^{-1} E_n - I) x_n \\ \qquad + (B^{-1} E_{n+1}^{(l)} - I) x_{n+1}^{(l+1)}. \end{cases}$$

下面给出格式 (I) 的几种变形.

1. 通常将 (13.1) 中的微分方程的两边乘上对角形常数矩阵 E 就可将 (13.1) 化成 (13.3) 的形式。 这时 $E(t, y) = E$. 假定 (13.4) 中的矩阵 B 仍如前所述为对角线常数矩阵. 令向量 $\tilde{z}(t) = B^{-1}u(t)$, $\tilde{x}_n = \frac{h}{2}\tilde{z}_n$. 易知

$$B^{-1}(E - B)x_n = B^{-1}(E - B)\frac{h}{2}z_n$$

$$= B^{-1}(E - B)\frac{h}{2}(E - B)^{-1}u_n$$

$$= B^{-1}\frac{h}{2}u_n = \frac{h}{2}\tilde{z}_n = \tilde{x}_n.$$

由此,格式 (I) 可以改写如下

(i) $y_{n+1}^{(0)} = e^{-B^{-1}h}y_n + 2e^{-B^{-1}h}\tilde{x}_n$,

(ii) $\tilde{x}_{n+1}^{(l+1)} = \left(I + \frac{2}{h}E\right)^{-1}\{g_{n+1}^{(l)} + B^{-1}(E - B)y_n + hB^{-1}g_n - \tilde{x}_n\}$,

(iii) $y_{n+1}^{(l+1)} = e^{-B^{-1}h}(y_n + \tilde{x}_n) + \tilde{x}_n^{(l+1)}$,

去掉"～"这一特殊记号，重新定义 $z(t) = B^{-1}u(t)$, $x_n = \frac{h}{2}z_n$, 即可得如下的格式 (II)

$$\text{II}\begin{cases}\text{(i)}\ y_{n+1}^{(0)} = e^{-B^{-1}h}y_n + 2e^{-B^{-1}h}x_n, \\ \text{(ii)}\ x_{n+1}^{(l+1)} = \left(I + \frac{2}{h}E\right)^{-1}\{g_{n+1}^{(l)} + B^{-1}(E - B)y_n + hB^{-1}g_n - x_n\}, \\ \text{(iii)}\ y_{n+1}^{(l+1)} = e^{-B^{-1}h}(y_n + x_n) + x_{n+1}^{(l+1)}.\end{cases}$$

2. 令 (13.4) 中的矩阵 $B \to \infty$ (即 B 的对角线上所有元素 $b_i \to \infty$)，则根据 (I) 中的定义有:

$$x_n = \frac{h}{2}z_n = \frac{h}{2}(E_n - B)^{-1}u_n$$

$$= \frac{h}{2}(E_n - B)^{-1}(By_n' + y_n),$$

当 $B\to\infty$ 时，$x_n=-\frac{h}{2}y'_n$，同时 $e^{-B^{-1}h}=I$. 由此，(I) 可改写为

$$\begin{cases}\text{(i)}\ \ y^{(0)}_{n+1}=y_n-2x_n,\\ \text{(ii)}\ \ x^{(l+1)}_{n+1}=\left(I+\frac{2}{h}E^{(l)}_{n+1}\right)^{-1}\{-g^{(l)}_{n+1}+y_n-x_n\},\\ \text{(iii)}\ \ y^{(l+1)}_{n+1}=y_n-x_n-x^{(l+1)}_{n+1}.\end{cases}$$

若重新定义 $x_n=\frac{h}{2}y'_n$，则 (I) 可简化为

$$\text{III}\begin{cases}\text{(i)}\ \ y^{(0)}_{n+1}=y_n+2x_n,\\ \text{(ii)}\ \ x^{(l+1)}_{n+1}=\left(I+\frac{2}{h}\ E^{(l)}_{n+1}\right)^{-1}\{g^{(l)}_{n+1}-y_n-x_n\},\\ \text{(iii)}\ \ y^{(l+1)}_{n+1}=y_n+x_n+x^{(l+1)}_{n+1}.\end{cases}$$

下面对格式 (III) 进行一些分析. 首先我们建立格式 (III) 与梯形公式的关系. 为此，将梯形公式的两端乘以常数 $(1-p)$，得

$$(1-p)y_{n+1}=(1-p)y_n+(1-p)\frac{h}{2}y'_n$$
$$+(1-p)\frac{h}{2}y'_{n+1}.$$

令 $x_n=\frac{h}{2}y'_n$，则有

$$y_{n+1}=y_n+x_n+p\left(\frac{1-p}{p}\cdot\frac{h}{2}f_{n+1}+y_{n+1}-y_n-x_n\right),$$

取 $p=\frac{h}{h+2\varepsilon}$，$\varepsilon$ 为小正数；则 $\frac{1-p}{p}=\frac{2\varepsilon}{h}$，故

$$y_{n+1}=y_n+x_n+p(\varepsilon f_{n+1}+y_{n+1}-y_n-x_n).$$

因而推出 $\frac{h}{2}y'_{n+1}=p(\varepsilon f_{n+1}+y_{n+1}-y_n-x_n)$，又令 $x_{n+1}=\frac{h}{2}y'_{n+1}$ 合并以上两式，得出格式

$$\text{(III}'\text{)}\begin{cases}x_{n+1}=p[\varepsilon f_{n+1}+y_{n+1}-y_n-x_n]\\ y_{n+1}=y_n+x_n+x_{n+1}\end{cases}$$

它恰好是格式（III）迭代到收敛时的形式．由此也可推出（III'）与梯形公式具有完全相同的截断误差．

将格式（III'）应用到试验方程

$$y' = \lambda y,\ \mathrm{Re}\lambda < 0.$$

格式（III'）可以写成下面的递推式

$$\begin{pmatrix} x_{n+1} \\ y_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & -p(1+\varepsilon\lambda) \\ -1 & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} -p & -p \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_n \\ y_n \end{pmatrix}. \tag{13.13}$$

容易算出，上面二矩阵的乘积的特征值 $\mu_1 = 0$,

$$\mu_2 = \frac{1-p+\varepsilon p\lambda}{1-p-\varepsilon p\lambda} = \frac{1+\dfrac{h}{2}\lambda}{1-\dfrac{h}{2}\lambda} < 1,\ 当\ 0<h<\infty\ 时, \tag{13.14}$$

并且 μ_2 和梯形公式应用于试验方程 $y' = \lambda y$ 的特征值完全一样．因此格式（III'）是 A 稳定的．

格式（III）实际上是基于简单迭代的一种方法．它可以改写成形式

$$y_{n+1}^{(l+1)} = y_n + x_n + p[\varepsilon f(t_{n+1}, y_{n+1}^{(l)}) + y_{n+1}^{(l)} - y_n - x_n], \tag{13.15}$$

因此迭代收敛性条件为

$$\left| p\left[\varepsilon \frac{\partial f}{\partial y}(t_{n+1}, y_{n+1}^{(l)}) + 1\right]\right| < 1. \tag{13.16}$$

若 $\dfrac{\partial f}{\partial y} = \lambda = \alpha + i\beta\ (\alpha < 0)$，考虑到 $p = \dfrac{h}{h+2\varepsilon} < 1\ (\varepsilon > 0)$，则只要有

$$(1+\varepsilon\alpha)^2 + \varepsilon^2\beta^2 = 1 + 2\varepsilon\alpha + (\alpha^2+\beta^2)\varepsilon^2 \leqslant 1, \tag{13.17}$$

则可导出条件(13.16)．选取

$$0 < \varepsilon \leqslant \frac{-2\alpha}{\alpha^2+\beta^2},$$

就可以作到这一点.

这里简单迭代的收敛性条件和步长 h 无关,即对 $0<h<\infty$ 迭代都收敛.

对于方程组的情形,若 Jacobi 矩阵 $\frac{\partial f}{\partial y}$ 的特征值为 $\lambda_k=\alpha_k+i\beta_k$, $k=1,\cdots,n$, 其中 $\alpha_k<0$, 则取

$$0<\varepsilon\leqslant\min_{1\leqslant k\leqslant n}\frac{-2\alpha_k}{\alpha_k^2+\beta_k^2},$$

即有同样的结论.

对于梯形公式,简单迭代要求满足收敛性条件 $\left\|\frac{h}{2}\frac{\partial f}{\partial y}\right\|<1$. 这是要通过步长 h 来实现的. 而对于迭代 (13.15) 的收敛性条件 (13.16) 是通过选取参数 ε 来实现的. 只要系统是稳定的,当 ε 取得充分小,条件 (13.16) 总是可以满足的.

对刚性问题来说,大都有这样一个特点,在其起始时一个短暂的时间间隔内,系统本身产生剧变(暂态过程). 精度要求 h 很小. 因而我们开始可以取 $\varepsilon=1$. 此时尽管 $\partial f/\partial y$ 很大. 但由于 $p=\frac{h}{h+2\varepsilon}=O(h)\ll 1$, (13.16) 仍然满足. 随着时间的推移, h 开始变大,迭代将要发散. 一旦发现此种情况,我们就缩小 ε.

假如我们要处理的方程组阶数不高,且 Jacobi 矩阵也很容易求出,则可以选

$$\varepsilon=E(t,y)=-\left(\frac{\partial f(t,y)}{\partial y}\right)^{-1},$$

此时格式 (III) 的 (ii) 成为

$$\begin{aligned}x_{n+1}^{(l+1)}&=\left(I+\frac{2}{h}E_{n+1}^{(l)}\right)^{-1}(g_{n+1}^{(l)}-y_n-x_n)\\&=\left(I-\frac{2}{h}\left(\frac{\partial f_{n+1}^{(l)}}{\partial y}\right)^{-1}\right)^{-1}\left[-\left(\frac{\partial f_{n+1}^{(l)}}{\partial y}\right)^{-1}\right.\\&\quad\left.\cdot f_{n+1}^{(l)}+y_{n+1}^{(l)}-y_n-x_n\right]\end{aligned}$$

$$= \left(-\frac{\partial f_{n+1}^{(l)}}{\partial y} + \frac{2}{h}\right)^{-1} \left[f_{n+1}^{(l)} - \left(\frac{\partial f_{n+1}^{(l)}}{\partial y}\right)(y_{n+1}^{(l)} - y_n - x_n) \right].$$

不难看出，对常系数线性方程组

$$y' = Ay$$

应用 $E(t, y)$ 的这种选取方法，经过一次迭代，即可得到准确的梯形公式. 事实上，令 $x_n = \frac{h}{2} Ay_n$，对任意选取的初始近似 $y_{n+1}^{(0)}$，有

$$x_{n+1}^{(1)} = \left(-A + \frac{2}{h}\right)^{-1} \left[Ay_{n+1}^{(0)} - A\left(y_{n+1}^{(0)} - y_n - \frac{h}{2} Ay_n\right)\right]$$

$$= \frac{h}{2} A \left(I - \frac{h}{2} A\right)^{-1} \left(I + \frac{h}{2} A\right) y_n,$$

$$y_{n+1}^{(1)} = y_n + x_n + x_{n+1}^{(1)}$$

$$= \left(I + \frac{h}{2} A\right) y_n + \frac{h}{2} A \left(I - \frac{h}{2} A\right)^{-1} \cdot \left(I + \frac{h}{2} A\right) y_n$$

$$= \left(I - \frac{h}{2} A + \frac{h}{2} A\right)\left(I - \frac{h}{2} A\right)^{-1} \cdot \left(I + \frac{h}{2} A\right) y_n$$

$$= \left(I - \frac{h}{2} A\right)^{-1} \left(I + \frac{h}{2} A\right) y_n$$

§2 光滑近似特解方法（SAPS）

这个方法是 Dahlquist 1968 年提出的. 但是方法的思想在 1952 年已给出来了. 设刚性方程组给成形式

$$\frac{dx}{dt} = -Ax + f(t, x, y), \qquad (13.18a)$$

$$\frac{dy}{dt} = g(t, x, y), \tag{13.18b}$$

其中 $x \in R^m$，$y \in R^n$，$t \in [0, T]$，A是分段常值的m阶矩阵，f和g是光滑函数．假定A的特征值比起f和g的 Lipschitz 常数要大得多．更精确地说，我们假定对于适当选取的与时间无关的ρ，模有不等式

$$\left\|\left(-A + \frac{\partial f}{\partial x}\right)^{-1}\right\|\left(\left\|\frac{\partial g}{\partial y}\right\| + 2 \times \left(\left\|\frac{\partial f}{\partial y}\right\|\left\|\frac{\partial g}{\partial x}\right\|\right)^{\frac{1}{2}}\right) \leqslant \rho < 1. \tag{13.18c}$$

光滑近似特解方法的思想是不去计算区间$(\tau, \tau + h)$上的(13.18.a)，(13.18.b)的完全解$\begin{pmatrix} x \\ y \end{pmatrix}$，而去确定近似解$\begin{pmatrix} p \\ q \end{pmatrix}$，其中$p(t)$是(13.18.$a$)的近似特解，而$q(t)$是(13.18.$b$)的相应的解．我们想构造$p(t)$和$q(t)$，使得有

$$p(t) = x(t) - e^{-At}(x(t) - p(t)) + \text{截断误差},$$

$$q(t) = y(t) - T(t)(x(t) - p(t)) + \text{截断误差},$$

其中$T(t)$是矩阵值函数．引进它的目的是想使得$q(t)$与x的快变部分近似无关．

由于项e^{-At}在边界层迅速衰减到可忽略的程度．当$t \geqslant \tau_{BL}$时，将在某种要求的精度内有

$$\begin{pmatrix} x \\ y \end{pmatrix} \approx \begin{pmatrix} p \\ q \end{pmatrix}.$$

当A的特征值的实部很小，e^{-At}表示为高频慢衰减振动，则即使对于大的t值，上述的近似仍不成立．我们将不考虑这种情形．

例 13.1　考虑

$$x' = -100x + 1 \qquad x(0) = x_0, \tag{13.19a}$$

$$y' = x - y \qquad y(0) = y_0. \tag{13.19b}$$

取$p(t) = \frac{1}{100}$，则

$$x(t)=p(t)+(x_0-1/100)e^{-100t},$$

将 $p(t)$ 代入 (13.19.b) 中的 x，得 $q(t)$ 的方程

$$q'=1/100-q \quad q(0)=y_0, \tag{13.20}$$

其解为

$$q(t)=\left(y_0-\frac{1}{100}\right)e^{-t}+\frac{1}{100}.$$

但 (13.19.b) 的 $y(t)$ 为

$$y(t)=1/100+\left(y_0+\frac{x_0-1/100}{99}-1/100\right)e^{-t}$$
$$-\frac{x_0-1/100}{99}e^{-100t}.$$

因此 $q(t)-y(t)$ 当 t 较大时仍比较大．但若将 (13.20) 中的初值换成

$$q(0)=y_0+(x_0-1/100)/99,$$

得到的 $q(t)$ 为

$$q(t)=1/100+\left(y_0+\frac{x_0-1/100}{99}-1/100\right)e^{-t},$$

则有

$$y(t)=q(t)-\frac{x_0-1/100}{99}e^{-100t},$$

所以这样选取的 $p(t)$、$q(t)$ 将能满足要求．

对于一般的常微分方程初值问题

$$\frac{dz}{dt}=h(t,z),\ z(0)=z_0,\ z\in R^{m+n}. \tag{13.21}$$

若 Jacobi 矩阵 $\dfrac{\partial H}{\partial z}$ 的特征值可以分成二组

$$\mathscr{A}=\{\lambda_1,\lambda_2,\cdots,\lambda_m\}$$

和

$$\mathscr{B}=\{\lambda_{m+1},\lambda_{m+2},\cdots,\lambda_{m+n}\}$$

使得当 $\lambda_i\in\mathscr{A}$ 和 $\lambda_j\in\mathscr{B}$ 时，有

$$\min|\lambda_i|\gg\max|\lambda_j|$$

并且 $\frac{\partial H}{\partial z}$ 随时间是慢变化，可以找到分段常值矩阵 M，且具有性质

$$\begin{pmatrix} x \\ y \end{pmatrix} = Mz \qquad x \in R^m,\ y \in R^n$$

和

$$\begin{pmatrix} -Ax + f(t, x, y) \\ g(t, x, y) \end{pmatrix} = MH(t, z)$$

大的特征值将局部地含在矩阵 A 中，于是 SAPS 方法就可以应用。这里我们不考虑如何选取变换矩阵 M，而假定已将(13.21)变换到(13.18)的情形。

由上面看到，为了使 SAPS 方法能有效地实施，必须解决下面的一些问题：

i) 如何去构造局部特解 $p(t)$；

ii) 如何计算矩阵 $T(t)$；

iii) 如何处理初值。

下面分别来解决这些问题。

首先考虑局部特解的构造问题。

如果 $V(t)$ 是向量值多项式，则利用未定系数法或有限次迭代

$$p^{(0)} = 0,\quad Ap^{(i)} = V - \frac{dp^{(i-1)}}{dt} \tag{13.22}$$

可确定方程

$$\frac{dp}{dt} = -Ap + V \tag{13.23}$$

的一个多项式解。由于这个多项式解是唯一的，它不能满足任意的初始条件。如果 $V(t)$ 不是多项式，则迭代过程(13.22)常常是发散的。但是若 $\|A^{-1}\|$ 很小，则用(13.22)迭代较少几次往往能提供方程(13.23)的好的近似解。例如，若 $p^{(i)}$ 是微分方程

$$\frac{dp^{(i)}}{dt} + Ap^{(i)} - V = -(-1)^i A^{-i} \frac{d^i V}{\partial t^i}$$

的精确解，并且其右端项具有一致小的模，则$p^{(i)}$将接近于 (13.23) 的某个解。这样的特解可以应用到 (13.18.a)。

令

$$t_k = k \cdot h,$$
$$p_k = p(t_k),\ q_k = q(t_k),$$
$$p^i = (p^i_k,\ p^i_{k+1/2},\ p^i_{k+1})^T,$$
$$q^i = (q^i_k,\ q^i_{k+1/2},\ q^i_{k+1})^T,$$

要求选取的步长 h 使它与 f 和 g 的 Lipschitz 常数的乘积是小的，而 $h/\|A^{-1}\|$ 常常可以非常大。在区间 $[t_k,\ t_{k+1}]$ 上的多项式 $p(t)$ 和 $q(t)$ 按下面的方式交替地应用两个过程迭代地构造。一个过程即是 SAPS 过程，它对于输入的 $p^i(t)$, $q^i(t)$ 在 $t = t_k,\ t_{k+1/2},\ t_{k+1}$ 上的值，计算

$$V(t) = f(t,\ p^i(t),\ q^i(t)),\quad t = t_k,\ t_{k+1/2},\ t_{k+1},$$

然后利用二次插值多项式得到向量值函数 $V(t)$，再由求解 (13.23) 得到新的 $p^{i+1}(t)$。SAPS 过程还将输入的 $q^i(t)$ 变换成 $q^{i+1/2}(t) = q^i(t) + \Delta q^i$，其中 Δq^i 将在后面来确定。另外一个过程是通常的积分过程。它利用 SAPS 过程得到的 $p^{i+1}(t)$ 和 $q^{i+1/2}(t)$ 积分组 (13.18.b)。例如应用不带平滑过程的梯形法则，记为 TRAP。我们可得下面的公式：令

$$g_j = g(t_j,\ p^{i+1}_j,\ q^{i+1/2}_j),$$

则 TRAP 过程给出

$$q^{i+1}_{k+1} = q^{i+1/2}_k + h(g_k + g_{k+1})/2,$$
$$q^{i+1}_{k+1/2} = (q^{i+1/2}_k + q^{i+1}_{k+1})/2 - h(g_{k+1} - g_k)/8,$$
$$q^{i+1}_k = q^{i+1/2}_k.$$

这样可以将一次迭代表示成

$$\begin{pmatrix} p^{i+1} \\ q^{i+1/2} \end{pmatrix} = \mathrm{SAPS}\begin{pmatrix} p^i \\ q^i \end{pmatrix}, \tag{13.24}$$

$$(q^{i+1}) = \mathrm{TRAP}\begin{pmatrix} p^{i+1} \\ q^{i+1/2} \end{pmatrix}, \tag{13.25}$$

称其为 SAPS-TRAP 迭代。[86] 对它的收敛性和收敛速度进行

了详细的研究. $p(t)$ 和 $q(t)$ 在 $[t_k, t_{k+1}]$ 上的第一次近似由前面区间上的多项式外插确定.

下面考虑矩阵 $T(t)$ 及初值 q_0 的确定问题.

由上面看到，我们用 SAPS 过程处理方程组 (13.18.a)，而用通常的方法处理 (13.18.b). SAPS 过程求得的 $p(t)$ 通常不满足初始条件. 若 (13.18.a) 和 (13.18.b) 之间的耦合性较小时，则这种初值不满足对 (13.18.b) 的影响较小. 若耦合性较大时，这种不满足性就不能不考虑了. 但是我们可以选取变换矩阵 $T(t)$，使得量 $y - Tx$ 相对于 y 来说与 x 有较小的耦合性. 这表示我们要寻找矩阵 $T(t)$，使得量

$$\begin{aligned} s(t) &= y(t) - T(t)x(t) - [q(t) - T(t)p(t)] \\ &= y(t) - [q(t) + T(t)(x(t) - p(t))] \end{aligned}$$

是小的并且是连续的. 换言之，我们要求矩阵 $T(t)$，使得 $q(t)$ 和 $x(t) - p(t)$ 之间的耦合性小. 为此，我们先考虑常系数的线性系统

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} -A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix},$$

要求寻找矩阵 T，使 $\hat{y} = y - Tx$ 与 x 不耦合. 即有

$$\begin{aligned} \begin{pmatrix} I & 0 \\ -T & I \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} x' \\ \hat{y}' \end{pmatrix} &= \begin{pmatrix} I & 0 \\ -T & I \end{pmatrix} \\ &\times \begin{pmatrix} -A & B \\ C & D \end{pmatrix} \begin{pmatrix} I & 0 \\ T & I \end{pmatrix} \begin{pmatrix} x \\ \hat{y} \end{pmatrix} \\ &= \begin{pmatrix} \hat{A} & \hat{B} \\ \hat{C} & \hat{D} \end{pmatrix} \begin{pmatrix} x \\ \hat{y} \end{pmatrix}, \end{aligned}$$

其中

$$\begin{aligned} \hat{A} &= -A + BT, \\ \hat{B} &= B, \\ \hat{C} &= -T(-A + BT) + C + DT, \\ \hat{D} &= D - TB. \end{aligned}$$

于是 T 是矩阵方程 $\hat{C} = 0$ 的解. 在方程 (13.18a), (13.18b)

的情况，首先将方程线性化，于是对应于 $\hat{C}=0$ 的方程为

$$T\left(-A+\frac{\partial f}{\partial p}+\frac{\partial f}{\partial q}T\right)=\frac{\partial g}{\partial p}+\frac{\partial g}{\partial q}T, \qquad (13.26)$$

为了求解方程(13.26)，可采用下面的迭代

$$T_0=0,$$

$$T_{i+1}\left(-A+\frac{\partial f}{\partial p}+\frac{\partial f}{\partial q}T_i\right)=\frac{\partial g}{\partial p}+\frac{\partial g}{\partial q}T_i. \qquad (13.27)$$

如果条件(13.18.c)满足，则迭代过程(13.27)是收敛的．特别当(13.18.c)中的 ρ 很小时，收敛速度是十分快的．

由于采用 SAPS 过程 $p(t)$ 和 $q(t)$ 在 $t=t_k$ 点是间断的．令 Δp_k, Δq_k 是在 $t=t_k$ 处的跳跃，即有

$$\Delta p_k=p(t_k+)-p(t_k-),$$
$$\Delta q_k=q(t_k+)-q(t_k-),$$

特别在 $t=0$ 时，$p(t_0-)=x_0$, $q(t_0-)=y_0$, 令 $R_p(t)$, $R_q(t)$ 是在节点之间的局部截断误差．令 $r=x-p$，则 r 是 x 的快变部分．现在可以将 $p(t)$ 和 $q(t)$ 写成为微分方程

$$\frac{dp}{dt}=-Ap+f(t,p,q)+R_p(t)$$
$$+\sum_{k=0}^{\infty}\Delta p_k\delta(t-t_k), \qquad (13.28a)$$

$$\frac{dq}{dt}=g(t,p,q)+Rq(t)$$
$$+\sum_{k=0}^{\infty}\Delta q_k\delta(t-t_k) \qquad (13.28b)$$

的精确解，其中 δ 是 Dirac 函数．对于 $r(t)$，我们得到

$$\frac{dr}{dt}=-Ar+f(t,x,y)-f(t,p,q)-Rp(t)$$
$$-\sum_{k=0}^{\infty}\Delta p_k\delta(t-t_k), \qquad (13.28c)$$

由 $s(t)$ 的连续性推出

$$\Delta q_k=T\Delta p_k, \qquad (13.29)$$

特别在 $t=0$ 时有 $(q_0-y_0)=T(p_0-x_0)$，因而

$$q_0=y_0+T(p_0-x_0).$$

现在考虑例 13.1，在 $t=0$ 时方程 (13.26) 为

$$T(-100)=1-T$$

即

$$T=-1/99.$$

方程 (13.29) 给出

$$q_0=y_0+T(p_0-x_0)=y_0-(1/100-x_0)/99.$$

由上面的叙述，可以将光滑近似特解方法的计算过程描述如下:

步 0　选取步长 h，令 $k=0$;

步 1　在 (t_k, p_k, q_k) 处计算矩阵 A^{-1} 和 T;

步 2　用 SAPS-TRAP 过程确定点 $(t_{k+1}, p_{k+1}, q_{k+1})$;

步 3　判定计算是否终止,若不停止,令 $k=k+1$，转步 1.

在实际计算时，矩阵 A，A^{-1}，T 不需要对每个 k 计算，可经过若干个点重新计算一次.

由于在 SAPS 过程中仅确定 (13.23)，(13.18a) 的稳态解，而将真解中的快速变化部分略去. 若在边界层外，由于这个略去的部分比较小,计算将是合理的. 但在边界层内,略去快速变化部分将会给计算带来很大的误差. 因而用 SAPS 过程计算将是不合理的. 另外在计算矩阵 T 时,对微分方程 (13.18a) 和 (13.18b) 进行了线性化. 这种处理在边界层内也是不合理的.由上述的理由,在边界层内必须采用别的计算方法. 例如采用梯形法则等传统的数值方法. 当在边界层外采用 SAPS 算法是比较有效的. 文 [93] 给出了这个算法的 Fortran 程序，并且给出了比较深入的理论和试验的分析.

§3　一类非线性方法

当将线性多步方法， Runge-Kutta 方法或其它许多方法应用到试验方程

$$y' = Ay$$

时，就得到离散变量 y_n 的线性差分方程．具有这种性质的数值求解初值问题．

$$y' = f(t, y), \quad y(t_0) = \eta \tag{13.30}$$

的方法称作线性方法．熟知，这些方法应用到刚性方程组时不是很满意的．为了具有适当的稳定性质，它们必须是隐式的．而为了求解，这些方法应用到刚性方程所产生的隐式差分方程组，不能使用简单迭代法，必须用某种 Newton 迭代．这就需要计算 Jacobi 矩阵的逆．如果 (13.30) 是一个大系统，求解将是很困难的，需要寻找避免计算 Jacobi 矩阵及其逆的数值方法．这一节给出构造显式非线性方法的一个途径，构造的方法具有处理刚性方程组的适当的稳定性，并且是显式的．

线性多步方法应用到 (13.30) 的基本思想是将 (13.30) 的解局部表示成多项式．由于刚性方程组的解中具有迅速衰减的指数分量，因此，若要求方法具有适当的稳定性（如 A 或 $A(\alpha)$ 稳定性），则上面的局部表示将限制线性多步方法的阶，并且要求方法是隐式的．为了能够看出这一点，考虑下面的简单的插值问题：给定函数 4^t 在 $t = 0, 1, 2, 3, \cdots$ 上的值，应用多项式插值求出 $t = \frac{1}{2}$ 时的值．容易看出最好的解答是在点 $x = 0$ 和 $x = 1$ 之间作线性内插．若用高次的多项式内插或外插将产生很坏的结果．因此用多项式来进行插值将限制插值多项式的次数，并且要求进行内插．但是，如果不用多项式，而用有理内插，上面插值问题中的困难就可以克服．由此想到不用局部多项式内插构造线性多步方法，而用局部有理插值来构造求解刚性方程的相应的方法．这样构造的方法是非线性方法．

§3.1　方法 I

我们首先给出一个初等的非线性方法．假定 (13.30) 是数初值问题，即 y 是一个标量，$y(t)$ 是它的真解．设在 $[t_n, t_{n+1}]$ 中

$y(t)$ 可局部地用有理函数

$$I(t) = A/(t + B) \tag{13.31}$$

来表示。如果 y_n 是 $y(t_n)$ 的近似，$f_n = f(t_n, y_n)$，$t_n = t_0 + nh$，由条件

$$y_n = I(t_n),\ y_{n+1} = I(t_{n+1}),\ f_n = I'(t_n)$$

消去 (13.31) 中的 A 和 B，得到方法

$$y_{n+1} - y_n = \frac{h y_n f_n}{y_n - h f_n}, \tag{13.32}$$

这个方法是与 Euler 公式类似。为了方法 (13.32) 的执行，对 $y(t)$ 必须加上限制

$$|y(t)| + |y'(t)| \neq 0,\ t \geqslant t_0, \tag{13.33}$$

若 (13.33) 成立，而 (13.32) 中的 $y_n - hf_n = 0$，则选取另外的 h 进行计算。

方法 (13.32) 也可以应用到方程组的情形，只要对每个分量进行计算。若 y 和 f 均是 m 维向量。$y = (y^1, y^2, \cdots, y^m)^T$，$f = (f^1, f^2, \cdots, f^m)^T$，对每个分量应用公式

$$y_{n+1}^i - y_n^i = \frac{h y_n^i f_n^i}{y_n^i - h f_n^i}. \tag{13.34}$$

方法 (13.34) 具有好的稳定性质。将 (13.34) 应用到试验方程 $y' = \lambda y$，$\mathrm{Re}\lambda < 0$，则有

$$y_{n+1}^i / y_n^i = 1/(1 - h\lambda)\quad i = 1, 2, \cdots, m. \tag{13.35}$$

因此，y_{n+1}^i / y_n^i 是 $e^{h\lambda}$ 的 (0, 1) Padé 近似，方法 (13.34) 是 A 稳定的。事实上是 L 稳定的。但是对于非线性方法，对试验方程的分析还不能直接应用到具有不同特征值的矩阵 A 的一般的线性方程组

$$y' = Ay, \tag{13.36}$$

其中 A 的特征值均在左半平面中，而对于线性方法是可以的。例如，作变换 $y = Hz$，其中矩阵 H 满足 $H^{-1}AH = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \cdots, \lambda_m)$。于是 (13.36) 变换成 $z' = \Lambda z$，即具有试验方程的形式。将 Euler 方法应用到方程(13.36)，得到 $y_{n+1} = (I + hA)y_n$。对

它应用变换 $y_n = Hz_n$,得到 $z_{n+1} = (I + h\Lambda)z_n$. 这恰好是 Euler 方法应用到变换后的方程组得到的结果。类似的结果对于非线性方法不成立.

下面我们只讨论数方程 (13.30) 和数试验方程 $y' = \lambda y$ 的情形.

类似于线性方法，我们来定义方法 (13.32) 的阶. 作非线性算子 $P[y(t); h]$:

$$P[y(t); h] = y(t + h) - y(t) - hy(t)y'(t)/[y(t) - hy'(t)], \tag{13.37}$$

其中 $y(t)$ 是 C^1 中的任何函数,对所有 t 满足 $|y(t)| + |y'(t)| \neq 0$. 如果有 $P[y(t); h] = O(h^{p+1})$, 则说方法是 p 阶的. 在 t_{n+1}, 方法 (13.32) 的局部截断误差 T_{n+1} 定义为

$$T_{n+1} = P[y(t_n); h],$$

其中 $y(t_n)$ 是初值问题的真解. 于是若 $y_n = y(t_n)$, 则由方法 (13.32)得到的 y_{n+1} 将有

$$y(t_{n+1}) - y_{n+1} = T_{n+1}.$$

显然，由方法 (13.32), 不管 t_n 是何值,只要 $y_n = 0$, 将得到 $y_{n+1} = 0$. 因此,当解通过零点时,方法就失败了. 可以通过考察方法的局部截断误差来说明这个现象. 将 $P[y(t_n); h]$ 在 t_n 处展开,我们得到

$$T_{n+1} = h^2\left[\frac{1}{2}y'' - (y')^2/y\right]_{t=t_n} + O(h^3), \tag{13.38}$$

这表示一般来说方法的阶为 1. 但当 $y(t_n) = 0$ 时, T_{n+1} 中 h^2 的系数无穷大. 这似乎表明当 $y_n = 0$ 时,方法的局部截断误差是无穷大. 但实际情况并不是这样. 当方程的真解通过零点后，数值解将沿着 t 轴. 因此用展开式 (13.38) 无法说明前面的现象. 下面我们将 $P[y(t); h]$ 的分子和分母分别展开,得到

$$T_{n+1} = \frac{[yy'' - 2(y')^2]h^2/2! + [yy''' - 3y'y'']h^3/3! + O(h^4)}{y - hy'}. \tag{13.39}$$

显然，如果 $y_n \neq 0$，$T_{n+1} = O(h^{p+1})$，$p \geqslant 1$. 而当 $y_n = 0$ 时，$T_{n+1} = O(h)$，即如果 $y_n \neq 0$，则方法的阶至少是 1. 而当 $y_n = 0$ 时，方法的阶为零，它是局部不相容的. 正是这种局部不相容性，导出当 $y_n = 0$ 时数值解所出现的现象. 由此也可以看出，对于非线性方法，"主"局部截断误差的概念将导出错误的结果. 因此这时应该将局部截断误差表示成(13.37)的形式. 另外，可以看到非线性方法的阶将依赖于初值问题，并且随着求解的过程而变化.

§ 3.2 方法 II

由于方法 (13.32) 当 $y_n = 0$ 时是不相容的，希望寻找不具有这种现象的别种方法. 方法 (13.32) 的这种局部不相容性可通过考察插值函数 $I(t) = A/(t + B)$ 的几何性质来说明. 这个函数只当 t 为无穷大时才可能是零，即在它的水平渐近线上. 利用这部分曲线进行外插，当然得到零解. 于是我们考虑应用局部插值函数

$$I(t) = \frac{At + B}{t + C} \tag{13.40}$$

对于有限的 t，它具有零点. 由条件

$$y_{n+j} = I(t_{n+j}),\ j = 0, 1, 2,\ f_{n+1} = I'(t_{n+1})$$

消去 (13.40) 中的 A、B、C，得到二步非线性方法

$$y_{n+2} - y_{n+1} = \frac{h(y_{n+1} - y_n)f_{n+1}}{2(y_{n+1} - y_n) - hf_{n+1}}, \tag{13.40'}$$

它的局部截断误差是

$$T_{n+2} = \left.\frac{\left[\frac{1}{3}y'y^{(3)} - \frac{1}{2}(y'')^2\right]h^3 + \left[\frac{1}{3}y'y^{(4)} - \frac{2}{3}y''y^{(3)}\right]h^4 + O(h^5)}{y' - \frac{1}{6}h^2y^{(3)} - \frac{1}{12}h^3y^{(4)} + O(h^4)}\right|_{t=t_n}. \tag{13.41}$$

由 (13.41) 推出，如果 $y'_n \neq 0$，则方法的阶至少是 2. 但是如果

$y'_n = 0$，$y''_n \neq 0$，方法的阶为零．因此，当解通过极大值或极小值点时，方法就失败．这个事实也可以由 (13.40′) 直接推出．因为不管 f 取何值,只要有 $y_{n+1} = y_n$，由 (13.40′) 得出 $y_{n+2}=y_{n+1}$。

为考察方法的稳定性，将方法 (13.40′) 应用到试验方程 $y' = \lambda y$，得到

$$y_{n+2} = y_{n+1} \frac{y_{n+1} - \left(1 + \frac{1}{2}\bar{h}\right)y_n}{\left(1 - \frac{1}{2}\bar{h}\right)y_{n+1} - y_n}, \quad \bar{h} = \lambda h. \tag{13.42}$$

令 $W_n = y_{n+1}/y_n$，给出

$$W_{n+1} = \frac{W_n - \left(1 + \frac{1}{2}\bar{h}\right)}{\left(1 - \frac{1}{2}\bar{h}\right)W_n - 1}. \tag{13.43}$$

应用下面的引理

引理 13.1 给定复数 a、b、c，满足 $a^2 \neq -bc$，令序列 W_n，$n = 0, 1, 2, \cdots$ 按下面的规则构成

$$W_{n+1} = \begin{cases} \dfrac{aW_n + b}{cW_n - a} & \text{如果 } W_n \text{ 是有限的，} \\[2ex] \dfrac{a}{c} & \text{如果 } W_n \text{ 是无限的，} \end{cases}$$

则

$$W_n = \begin{cases} W_0 \text{ 如果 } n \text{ 是偶数，} \\ W_1 \text{ 如果 } n \text{ 是奇数。} \end{cases}$$

我们有

$$y_n/y_0 = \prod_{i=0}^{n-1} W_i = \begin{cases} (W_0W_1)^{n/2}, & \text{如果 } n \text{ 是偶数，} \\ (W_0W_1)^{(n-1)/2}W_0, & \text{如果 } n \text{ 是奇数。} \end{cases}$$

这样,当 $n \to \infty$ 时，$y_n \to 0$ 的主要条件为 $|W_0W_1| < 1$，即 $|y_2/y_0| < 1$．特别,如果附加的起始值 y_1 是由梯形法则得到的，$W_0 = y_1/y_0 = \left(1 + \frac{1}{2}\bar{h}\right) \Big/ \left(1 - \frac{1}{2}\bar{h}\right)$．将其代入 (13.43)，我们看到

W_1 也是 $\left(1+\frac{1}{2}\bar{h}\right)\Big/\left(1-\frac{1}{2}\bar{h}\right)$. 于是由引理 13.1，对所有 n，有

$$\frac{y_{n+1}}{y_n}=\frac{1+\frac{1}{2}\bar{h}}{1-\frac{1}{2}\bar{h}},$$

因此得到的方法是 A 稳定的.

现在我们来讨论方法 (13.40′)，当 $y'_n=0$，$y''_n\neq 0$ 的局部不相容性. 可以用讨论插值函数 $I(t)$ 的几何图形来解释这种现象. 这个函数只在无穷大的 t 处才可能有斜率为零，即在它的水平渐近线上. 因此由它外插给出 $y_{n+2}=y_{n+1}$.

§ 3.3 方法 III

为了克服方法 II 中的缺点，在方法中引进解的曲率的信息. 为此，仍使用插值函数 (13.40) 但要求它满足条件

$$y_{n+j}=I(t_{n+j}),\ j=0,1,\ f_n=I'(t_n),\ f'_n=I''(t_n),$$

其中 $f'_n=f'(t_n,y_n)$ 是由对 (13.30) 中的微分方程进行微分所得到的对 t 的全导数. 我们得到单步非线性方法

$$y_{n+1}-y_n=\frac{2hf_n^2}{2f_n-hf'_n},\quad |y'(t)|+|y''(t)|\neq 0. \tag{13.44}$$

它的局部截断误差是

$$T_{n+1}=\left.\frac{\left[-\frac{1}{2}(y'')^2+\frac{1}{3}y'y'''\right]h^3+O(h^4)}{2y'-hy''}\right|_{t=t_n} \tag{13.45}$$

如果 $y'_n\neq 0$，则方法的阶至少为 2，而如果 $y'_n=0$，则方法的阶为 1. 因此，这个方法避免了局部不相容性. 注意，若将 (13.45) 按通常的方式展开，得到

$$T_{n+1}=h^3\left[\frac{1}{6}y'''-\frac{1}{4}\frac{(y'')^2}{y'}\right]_{t=t_n}+O(h^4).$$

当 $y_n' = 0$ 时，这种表示式无效。

将方法应用到试验方程 $y' = \lambda y$，得到

$$\frac{y_{n+1}}{y_n} = \frac{1+\frac{1}{2}\bar{h}}{1-\frac{1}{2}\bar{h}}$$

因此方法是 A 稳定的。

§3.4 方法IV

应用插值函数

$$I(t) = \frac{At+B}{t^2+Ct+D}, \tag{13.46}$$

并要求满足条件

$$y_{n+j} = I(t_{n+j}) \quad j = 0, 1, 2, \quad f_{n+1} = I'(t_{n+1}), \quad f'_{n+1} = I''(t_{n+1})$$

得到二步非线性方法

$$y_{n+2} = \frac{4y_{n+1}^2(y_{n+1}-y_n)-4hy_{n+1}y_nf_{n+1}-h^2(2f_{n+1}^2-y_{n+1}f'_{n+1})y_n}{4y_{n+1}(y_{n+1}-y_n)-4hy_{n+1}f_{n+1}+h^2[2f_{n+1}^2-(y_{n+1}-2y_n)f'_{n+1}]}, \tag{13.47}$$

$$|y(t)| + |y'(t)| + |y''(t)| \doteq 0.$$

由这个方法的局部截断误差的表示式(太繁复，省略)可以导出方法的阶 p 为

$$\begin{aligned} &p \geqslant 3 \quad \text{一般情形} \\ &p \geqslant 2 \quad \text{如果}\ y_n' = y_n'' = 0, \\ &p = 1 \quad \text{如果}\ y_n = y_n' = 0. \end{aligned} \tag{13.48}$$

将方法 (13.47) 应用到试验方程 $y' = \lambda y$，并令 $W_n = y_{n+1}/y_n$，作一些运算后，我们得到

$$W_{n+1} = \frac{W_n - \left(1+\frac{1}{2}\bar{h}\right)^2}{\left(1-\frac{1}{2}\bar{h}\right)W_n - \left(1-\frac{1}{2}\bar{h}^2\right)}. \tag{13.49}$$

引理 13.1 已不能应用到这个方程．需要研究非线性差分方程

$$W_{n+1}=\frac{aW_n+b}{cW_n+d}$$

的解的性质．显然，这个方程存在两个常数解即 $W_n=\hat{W}$，其中 $\hat{W}$ 是特征多项式 $cW^2+(d-a)W-b$ 的两个根中的一个．下面的引理及其推论将提供研究 (13.49) 的解的工具．

引理 13.2 设 a、b、c、d 均是复数，且 $ad\neq bc$，$c\neq 0$ 和 $b\neq 0$．令序列 W_n，$n=0,1,2,\cdots$ 由下面的规则构成

$$W_{n+1}=\begin{cases}\dfrac{aW_n+b}{cW_n+d} & \text{如果 } W_n \text{ 是有限的,}\\[2ex] \dfrac{a}{c} & \text{如果 } W_n \text{ 是无限的.}\end{cases}$$

于是 $W_n=\hat{W}(1+\varepsilon_n)$

$$\varepsilon_n=\frac{A^n\varepsilon_0}{C\left[\sum\limits_{j=0}^{n-1}A^{n-j-1}D^j\right]\varepsilon_0+D^n},$$

其中 $\hat{W}$ 是 $cW^2+(d-a)W-b=0$ 的根，$A=a-c\hat{W}$，$C=c\hat{W}$ 和$D=d+c\hat{W}$.

推论 1 对于 $n=0,1,\cdots,m-1$，W_n 是有限的而 W_m 是无限的充分必要条件是起始值 $W_0(=\hat{W}(1+\varepsilon_0))$ 有

$$\varepsilon_0=\frac{-D^m}{C\sum\limits_{j=0}^{m-1}A^{m-j-1}D^j}.$$

推论 2 设特征多项式 $cW^2+(d-a)W-b$ 有重根 $\hat{W}$（即 $A=D$），则

(i) 对于最多一个 n 值，W_n 是无限的；

(ii) 如果对于某个正整数 m，$\varepsilon_0=-A/mC$，则 W_m 是无限的，而对 $n\neq m$，W_n 是有限的；

(iii) 如果对任意的正整数 m，$\varepsilon_0\neq -A/mC$，则对所有的 n，W_n 是有限的，和

(iv) $$W_n = \hat{W}\left[1 + \frac{A\varepsilon_0}{A + nC\varepsilon_0}\right],$$

差分方程 (13.49) 的特征多项式恰好具有重根，它为

$$\hat{W} = \frac{1 + \frac{1}{2}\bar{h}}{1 - \frac{1}{2}\bar{h}}, \tag{13.50}$$

并且 $A = \frac{1}{4}\bar{h}^2$, $C = 1 - \frac{1}{4}\bar{h}^2$.

现在分两种情形来讨论.

情形 (i)，假定不存在正整数 m，使得有 $\varepsilon_0 = -A/mC$，则 W_n 总是有限的. 令 $\bar{h}$ 固定，且 $\mathrm{Re}\bar{h} < 0$，则 A, C, ε_0 和 $\hat{W}$ 均是固定的. 由 (13.50) 推得

$$|\hat{W}| \leqslant K < 1,$$

再由当 $n \to \infty$ 时

$$1 + \frac{A\varepsilon_0}{A + nC\varepsilon_0} \to 1,$$

存在正整数 N，使得对所有 $n > N$，有

$$\left|1 + \frac{A\varepsilon_0}{A + nC\varepsilon_0}\right| < \frac{K}{|\hat{W}|}.$$

于是由推论 2 的 (iv)，有

$$\left|\frac{y_{N+n}}{y_N}\right| = \left|\prod_{i=N}^{N+n-1} W_i\right| \leqslant \prod_{i=N}^{N+n-1} |W_i| < K^{n-1}$$

因此，由于 y_N 一定是有限的，当 $n \to \infty$ 时 $y_{N+n} \to 0$.

情形 (ii)，假定存在一个正整数 m，使有 $\varepsilon_0 = -A/mC$，由推论 2 的 (ii)，若将情形 (i) 中的 N 换成 $\tilde{N} = \max(N, m + 1)$，则其中的讨论也成立.

综合情形 (i) 和 (ii) 可知方法 (13.47) 是 A 稳定的. 由于当 $\bar{h} \to -\infty$ 时，$\hat{W} \to -1$，$W_n \to -\left(1 + \frac{\varepsilon_0}{1 - n\varepsilon_0}\right)$. 因此方法 (13.47) 不是 L 稳定的.

§ 3.5 方法 V

仍应用插值函数 (13.46)，并要求满足条件 $y_{n+j}=I(t_{n+j})$，$j=0,1$，$f_n=I'(t_n)$，$f'_n=I''(t_n)$，$f_n^{(2)}=I^{(3)}(t_n)$．得到方法

$$y_{n+1}=y_n+hf_n+\frac{1}{2}h^2f'_n+\frac{1}{6}h^3\frac{P_nf''_n+hR_nf'_n}{P_n+hQ_n-\frac{1}{3}h^2R_n},$$

$$|y(t)|+|y'(t)|+|y''(t)|\neq 0, \qquad (13.51)$$

$$|y'(t)|+|y''(t)|+|y'''(t)|\neq 0,$$

其中

$$P_n=y_nf'_n-2f_n^2,\quad Q_n=f_nf'_n-\frac{1}{3}y_nf''_n,$$

$$R_n=\frac{3}{2}(f'_n)^2-f_nf_n^{(2)}.$$

这个方法的局部截断误差非常复杂，这里省略．方法 (13.51) 的阶恰好与方法 (13.47) 的相同，并由 (13.48) 给出．

将方法 (13.51) 应用到试验方程 $y'=\lambda y$，得到

$$\frac{y_{n+1}}{y_n}=\frac{1+\frac{1}{3}\bar{h}}{1-\frac{2}{3}\bar{h}+\frac{1}{6}\bar{h}^2},$$

它恰好是 $e^{\bar{h}}$ 的(1,2) Pade′ 近似．因此方法是 A 稳定的．事实上还是 L 稳定的．

在 [68] 中 Lambert 还叙述了 Runge-Kutta 型算法．

方法 VI (Runge-Kutta 型)

$$y_{n+1}-y_n=hf\left(t_n+\frac{1}{2}h,\ y_n+\frac{1}{2}\frac{hy_nf_n}{y_n-\frac{1}{2}hf_n}\right),$$

$$|y(t)|+|y'(t)|\neq 0,$$

如果 $y_n\neq 0$，则 $p\geqslant 2$．如果 $y_n=0$，则 $p=1$．方法是 A 稳定的．

方法 VII (Runge-Kutta 型)

$$y_{n+1}-y_n=\frac{hy_n(c_1K_1+c_2K_2)}{y_n-h(d_1K_1+d_2K_2)},$$

$$K_1=f(t_n, y_n),$$

$$K_2=f\left(t_n+ah, y_n+\frac{ahK_1y_n}{y_n-ahK_1}\right),$$

$$|y(t)|+|y'(t)|\neq 0.$$

若 $c_1=1-1/2a$, $c_2=1/2a$, $d_1=-d_2=-c_1$, 则如果 $y_n\neq 0$, $p\geqslant 2$, 如果 $y_n=0$, $p=0$, 对于 $a>\frac{1}{2}$, 方法是 L 稳定的.

由上面列举的方法的例子,所构造的方法在解的特殊点上,阶将减低. 另外对一般线性方程 $y'=Ay$ 的稳定性分析,尚无结果.

§ 4 矩阵分解方法(系统方法)

许多求解常微分方程初值问题的计算方法是基于经典的 Taylor 展开公式. 将这展开推广, 得到矩阵分解. 这一节给出的矩阵分解的数值方法对于常系数的线性常微分方程组是精确的.

§ 4.1 线性系统的数值求解方法

如果可以找到初值问题的解析解, 则利用它我们可以建立差分方程组,使得以任意的离散步长这个解将精确地满足这个组.

首先考虑非齐次线性方程组

$$x'(t)=Ax(t)+b, \ x(0)=x_0, \ x(t)\in R^m, \ t\in[0, T], \tag{13.52}$$

(13.52) 的解有形式

$$x(t)=\exp(At)x_0+\int_0^t \exp(A\tau)d\tau, b \tag{13.53}$$

矩阵指数函数 $\exp(At)$ 可以表成一致收敛的矩阵级数

$$\exp(At) = \sum_{s=0}^{\infty} \frac{A^s t^s}{s!}, \quad A^0 = E, \tag{13.54}$$

其中 E 是单位矩阵．相应地，$\exp(At)$ 的积分有表示式

$$\int_0^t \exp(At)dt = \sum_{s=0}^{\infty} \frac{A^s t^{s+1}}{(s+1)!}, \tag{13.55}$$

但是对于充分大的 t 值，直接应用 (13.54) 和 (13.55) 来计算 $\exp(At)$ 及其积分是不合理的，因为为了达到需要的精度，必须取非常多的项．这对于刚性方程更严重．

不用 (13.54) 和 (13.55)，我们直接由公式 (13.53) 构造一个差分方程组，一步接一步计算 (13.52) 的解．选取 H，对于 $t+H$，将 (13.53) 写成

$$x(t+H) = \exp(At + AH)x_0 + \int_0^{t+H} \exp(A\tau)d\tau b, \tag{13.56}$$

用 $\exp(AH)$ 乘 (13.53)，并代人 (13.56)，得到

$$x(t+H) = \exp(AH)x(t) + \int_0^H \exp(A\tau)d\tau b.$$

记 $t_n = nH$，$n = 0, 1, 2, \cdots$，$x(t_{n+1}) = x_{n+1}$，$x(t_n) = x_n$，则得到差分方程

$$x_{n+1} = \exp(AH)x_n + \int_0^H \exp(A\tau)d\tau b, \quad x_n|_{n=0} = x_0, \tag{13.57}$$

应用 (13.57) 时，可先将矩阵 $\exp(AH)$，$\int_0^H \exp(A\tau)d\tau$，或矩阵 $\exp(AH)$ 和向量 $\int_0^H \exp(A\tau)d\tau b$ 计算好，它们不依赖于 t_n．为了计算它们，可应用递推公式．设 $h = H/2^N$，其中 N 是充分大的自然数．计算 $\exp(Ah)$，于是应用公式

$$\begin{gathered} \exp(2Ah) = \exp(Ah)\exp(Ah), \\ \cdots\cdots \\ \exp(2^N Ah) = \exp(2^{N-1}Ah)\exp(2^{N-1}Ah), \end{gathered} \tag{13.58}$$

计算 $\exp(AH) = \exp(2^N Ah)$． 公式 (13.58) 可写成矩阵递推关系式

$$\varphi_{k+1} = \varphi_k^2, \tag{13.59}$$

这里

$$\varphi_k = \exp(2^k Ah), \ k = 0, 1, 2, \cdots, N, \tag{13.60}$$

显然 $\varphi_N = \exp(AH)$. 这样，我们可以选取任意小的 h，使得 $\exp(Ah)$ 的展开式 (13.54) 用很少的几项就能达到任意要求的精度. 取

$$\varphi_0 = \exp(Ah) \simeq E + Ah + \cdots + \frac{A^\nu h^\nu}{\nu!}, \tag{13.61}$$

然后应用递推关系式 (13.59) 确定 $\exp(AH)$. 现在来建立计算矩阵

$$\Phi_N = \int_0^H \exp(A\tau)d\tau$$

的递推关系式. 先证明下面的公式

$$\int_0^{2h} \exp(A\tau)d\tau = (E + \exp(Ah)) \int_0^h \exp(A\tau)d\tau. \tag{13.62}$$

事实上，有

$$\begin{aligned}
&(E + \exp(Ah)) \int_0^h \exp(A\tau)d\tau \\
&= \int_0^h \exp(A\tau)d\tau + \int_0^h \exp(Ah + A\tau)d\tau \\
&= \int_0^h \exp(A\tau)d\tau + \int_h^{2h} \exp(A\tau)d\tau \\
&= \int_0^{2h} \exp(A\tau)d\tau.
\end{aligned} \tag{13.63}$$

(13.62) 可以写成形式

$$\Phi_{k+1} = \Phi_k(E + \varphi_k), \tag{13.64}$$

其中

$$\Phi_k = \int_0^{2^k h} \exp(A\tau)d\tau, \ k = 0, 1, \cdots, N, \tag{13.65}$$

而 φ_k 仍由 (13.60) 给出. 由 (13.54) 和 (13.55) 得出等式

$$\exp(2^k Ah) = \varphi_k = E + A\Phi_k = E + A \int_0^{2^k h} \exp(A\tau)d\tau. \tag{13.66}$$

利用这个等式可将关系式 (13.64) 变换成

$$\Phi_{k+1} = \Phi_k(2E + A\Phi_k), \tag{13.67}$$

这样的 Φ_k 的计算与 φ_k 无关.

为了减少求解 (13.52) 时的运算次数,我们不用 Φ_N, 而直接建立向量

$$g_k = \int_0^{2^k h} \exp(A\tau)d\tau b = \Phi_k b,\ k = 0, 1, \cdots, N \tag{13.68}$$

的递推关系式. 因为矩阵 Φ_k 和 $E + \varphi_k$ 对乘法是可交换的,由 (13.64) 和 (13.68), 有

$$g_{k+1} = (E + \varphi_k)g_k, \tag{13.69}$$

关系式 (13.69) 应该与 (13.59) 一起递推.

对于充分小的 h, 类似于 (13.61), 为了计算 Φ_0, 只要用展开式(13.55) 的很少几项,

$$\Phi = \int_0^h \exp(A\tau)d\tau \simeq h\left(E + \frac{Ah}{2} + \cdots + \frac{A^{\nu-1}h^{\nu-1}}{\nu!}\right). \tag{13.70}$$

由差分方程 (13.57) 及递推关系式 (13.59) 和 (13.69) 可得到 (13.52) 的解的递推式

$$z(nH + H) = \bar{\varphi}_N z(nH) + \bar{g}_N,\ n = 0, 1, \cdots, H = 2^N h \tag{13.71}$$

为了看出格式 (13.71) 的优越性,由四项 Taylor 展开式, 构造区间 $[0,T]$ 上的差分方程

$$\begin{aligned} z(nh + h) &= \left(E + Ah + \frac{A^2h^2}{2} + \frac{A^3h^3}{6} + \frac{A^4h^4}{24}\right)z(nh) \\ &\quad + h\left(E + \frac{Ah}{2} + \frac{A^2h^2}{6} + \frac{A^3h^3}{24}\right)b \\ &= \bar{\varphi}_0 z(nh) + \bar{\Phi}_0 b = \bar{\varphi}_0 z(nh) + \bar{g}_0, \end{aligned} \tag{13.72}$$

选取 h 使其能保证适当的精度. 格式 (13.72) 相应于四阶 Runge-Kutta 方法.对于刚性方程, h 取得很小,因而 T/h 很大(假定 T/h 是整数). 应用格式 (13.72) 需要 T/h 个矩阵和向量的乘法. 而应用格式 (13.71) 仅需要

$$T/H = 2^{-N}T/h$$

个矩阵与向量的乘法，且利用 (13.59) 和 (13.69) 计算 $\bar{\varphi}_N$ 和 $\bar{g}_N$ 的 $N(m+1)$ 个这样的运算，其中 m 是 (13.52) 的维数．这里为了简单起见，不考虑 (13.71) 和 (13.72) 中向量加法及计算 φ_0 和 g_0 中的运算．为了方便，取

$$T/h = 2^{17},\ m+1 = 2^6,\ N = 8,$$

则格式 (13.72) 的运算量是 (13.71) 的 2^7 倍．

格式 (13.71) 中的离散步长 $H = 2^N h$ 可以任意大，而为了保证精度，h 应取得适当小．下面再详细说明用格式 (13.71) 的计算过程．

1.按给定的步长 H 选取数 N 使得量 $h = H/2^N$ 充分小．对于刚性方程 (13.52) 取

$$h < \frac{1}{\|A\|}, \tag{13.73}$$

其中 $\|A\|$ 是所取的矩阵模．如果 (13.52) 不是刚性的，则 h 选成满足下式

$$h \ll \frac{1}{\|A\|}. \tag{13.74}$$

2. 利用公式 (13.70) 计算 Φ_0，其中数 ν 取得不超过 4，对刚性方程取 $\nu = 2$ 就可以了

3. 由 Φ_0，按公式 (13.66) 和 (13.68) 计算 φ_0 和 g_0．

4. 由 φ_0 和 g_0，按关系式 (13.59) 和 (13.69) 计算 $\bar{\varphi}_k$ 和 $\bar{g}_k$．公式 (13.59) 和 (13.69) 应用 N 次．

5. 利用 $\bar{\varphi}_N$ 和 $\bar{g}_N$，按格式 (13.71) 计算 (13.52) 的近似解．

6. 检验得到的解的精度．可以将 h 缩小一倍，进行计算，将计算结果进行比较．

实际计算表明应用(13.59)，(13.64) 和 (13.69) 求解微分方程组的算法对计算误差的敏感性比应用公式 (13.67) 的算法要小．这现象是因为 (13.67) 中应用了与矩阵 A 的乘法，而比起 $\bar{\varphi}_k$ 来矩阵 A 具有相当大的元素．因此对于高阶的矩阵及太大的 N 应用 (13.67) 是不合理的．

下面考虑求解形式为

$$\frac{dx}{dt}=Ax+g(t),\ g(t)\in C^d[0,T],\ x(0)=x_0 \tag{13.75}$$

的方程的方法。类似 (13.57)，方程 (13.75) 的解可写成

$$x(t+H)=\exp(AH)x(t)+\int_0^H \exp(AH-A\tau)g(t+\tau)d\tau. \tag{13.76}$$

引进离散值 $t_n=nH$. 假定对于选定的H和每一个值 t_n，向量值函数 $g(t_n+\tau)$ 可以以充分高的精度表示成幂次多项式

$$g(t_n+\tau)=\sum_{\gamma=0}^{\nu} M_\gamma(t_n,H)\frac{\tau^\gamma}{\gamma!}+r_\nu(g(t_n),\tau,H), \tag{13.77}$$

其中 $r_\nu(g(t_n),\tau,H)$ 是展开式的余项.

将 (13.77) 代入 (13.76)，得到

$$\begin{aligned}x(t_n+H)&=\exp(AH)x(t_n)\\&\quad+\sum_{\gamma=0}^{\nu}\int_0^H \exp(AH-A\tau)\frac{\tau^\gamma}{\gamma!}d\tau M_\gamma(t_n,H)\\&\quad+\int_0^H \exp(AH-A\tau)r_\nu(g(t_n),\tau,H)d\tau.\end{aligned} \tag{13.78}$$

记

$$\Phi_{N,\gamma+1}=\int_0^H \exp(AH-A\tau)\frac{\tau^\gamma}{\gamma!}d\tau,\ \gamma=0,1,\cdots,$$

$$\nu\leqslant d-1,$$

略去公式 (13.78) 中的余项

$$\int_0^H \exp(AH-A\tau)r_\nu(g(t_n),\tau,H)d\tau,$$

得到近似差分方程

$$z_{n+1}=\exp(AH)z_n+\sum_{\gamma=0}^{\nu}\Phi_{N,\gamma+1}M_\gamma(t_n,H). \tag{13.79}$$

为了用 (13.79) 求解，必须得到矩阵 $\Phi_{N,\gamma+1}$. 前面叙述的方

法给出 $\Phi_N = \Phi_{N,1}$. 记

$$\Phi_{k,\gamma+1} = \int_0^{2^k h} \exp(2^k Ah - A\tau)\frac{\tau^\gamma}{\gamma!}d\tau, \quad \gamma = 0, 1, \cdots, \nu,$$

$$k = 0, 1, \cdots, N; \; H = 2^N h, \tag{13.80}$$

应用关于卷积的公式

$$\int_0^t \varphi(t-\tau)\psi(\tau)d\tau = \int_0^t \varphi(\tau)\psi(t-\tau)d\tau,$$

可以将(13.80)写成形式

$$\int_0^{2^k h} \exp(2^k Ah - A\tau)\frac{\tau^\gamma}{\gamma!}d\tau = \int_0^{2^k h} \exp(A\tau)\frac{(2^k h - \tau)^\gamma}{\gamma!}d\tau$$

$$= \int_0^{2^k h} d\tau_1 \int_0^{\tau_1} d\tau_2 \cdots \int_0^{\tau_\gamma} \exp(A\tau_{\gamma+1})d\tau_{\gamma+1}. \tag{13.81}$$

矩阵 $\Phi_{k,\gamma+1}$ 是矩阵指数的 $(\gamma+1)$ 重积分. 对公式(13.54)积分 $(\gamma+1)$ 次,得到矩阵指数的 $\gamma+1$ 重积分的幂表示

$$\Phi_{k,\gamma+1} = \int_0^{2^k h} \exp(A2^k h - A\tau)\frac{\tau^\gamma}{\gamma!}d\tau$$

$$= \sum_{\alpha=0}^{\infty} \frac{(2^k h)^{\gamma+1+\alpha}A^\alpha}{(\gamma+1+\alpha)!}, \tag{13.82}$$

利用级数(13.82),由归纳法,容易证明公式

$$\Phi_{k,\gamma} = A\Phi_{k,\gamma+1} + \frac{(2^k h)^\gamma}{\gamma!}E; \; \gamma = 0, 1, \cdots, \nu;$$

$$k = 0, 1, \cdots N. \tag{13.83}$$

若记 $\varphi_k = \Phi_{k,0}$, 则当 $\gamma = 0$ 由(13.83)推得(13.66). 利用(13.66)和(13.59),可以得到(13.67).按照(13.83),公式(13.66)改写成

$$\Phi_{k,0} = A\Phi_{k,1} + E. \tag{13.84}$$

利用 $k+1$ 时的(13.84),应用(13.59),再用(13.84),有

$$(A\Phi_{k,1} + E)(A\Phi_{k,1} + E) = A\Phi_{k+1,1} + E \tag{13.85}$$

或者

$$2A\Phi_{k,1} + A\Phi_{k,1}A\Phi_{k,1} = A\Phi_{k+1,1}. \tag{13.86}$$

因为 (13.86) 对于任意矩阵 A 均成立，则由 (13.86) 推出 (13.67).

现在来求 $\Phi_{k,2}$ 的递推关系式. 利用 $\gamma = 1$ 时的公式 (13.83), 有

$$\Phi_{k,1} = A\Phi_{k,2} + 2^k hE. \tag{13.87}$$

应用由 $\Phi_{k,1}$ 得到 $\Phi_{k+1,1}$ 的公式 (13.67), 那么由 (13.87) 得到

$$\begin{aligned}(A\Phi_{k,2} + 2^k hE)(2E + A^2\Phi_{k,2} + 2^k hA) \\ = A\Phi_{k+1,2} + 2^{k+1}hE,\end{aligned} \tag{13.88}$$

(13.88) 对任意矩阵 A 均成立，推得递推关系式

$$\Phi_{k+1,2} = \Phi_{k,2}(2E + 2^{k+1}hA + A^2\Phi_{k,2}) + 2^{2k}h^2E, \quad k = 0, 1, \cdots, N-1. \tag{13.89}$$

为了减小计算 $\Phi_{N,2}$ 时舍入误差的影响，最好将 (13.89) 改写成不显含矩阵 A 和 A^2 的公式. 为此，应用 (13.87) 和 (13.84), 将 (13.89) 变换成形式

$$\begin{aligned}\Phi_{k+1,2} &= \Phi_{k,2}(2E + A\Phi_{k,1} + 2^k hA) + 2^{2k}h^2E \\ &= \Phi_{k,2}(2E + A\Phi_{k,1}) + (\Phi_{k,1} - 2^k hE)2^k h + 2^{2k}h^2E \\ &= \Phi_{k,2}(E + \Phi_{k,0}) + 2^k h\Phi_{k,1}, \ k = 0, 1, \cdots, N-1.\end{aligned} \tag{13.90}$$

这样，在按递推公式

$$\Phi_{k+1,2} = \Phi_{k,2}(E + \Phi_{k,0}) + 2^k h\Phi_{k,1} \tag{13.91}$$

递推地计算 $\Phi_{N,2}$ 时，必须同时按公式 (13.59) 和 (13.64) 计算 $\Phi_{N,0}$ 和 $\Phi_{N,1}$.

应用 (13.83), 由数学归纳法可以得到确定 $\Phi_{N,\gamma+1}$ 的递推公式

$$\begin{aligned}\Phi_{k+1,\gamma+1} = \Phi_{k,\gamma+1}\left(2\sum_{\alpha=0}^{\gamma}\frac{A^\alpha(2^k h)^\alpha}{\alpha!} + A^{\gamma+1}\Phi_{k,\gamma+1}\right) \\ + \sum_{\beta=1}^{\gamma}\frac{(2^k h)^\beta}{\beta!}\sum_{\alpha=1}^{\beta}\frac{(2^k h)^{\gamma+1-\alpha}A^{\beta-\alpha}}{(\gamma+1-\alpha)!}, \ k = 0, \cdots, N-1.\end{aligned} \tag{13.92}$$

在实际计算 $\Phi_{N,\gamma+1}$ 时，最好从公式 (13.92) 中消去矩阵 A 及其幂次，同时计算 $\Phi_{N,0}, \Phi_{N,1}, \cdots, \Phi_{N,\gamma_0}$. 为了计算初始矩阵 $\Phi_{0,\gamma+1}$,

可以利用和

$$\sum_{\alpha=0}^{\Theta} \frac{h^{\gamma+1+\alpha}A^{\alpha}}{(\gamma+1+\alpha)!} = \Phi_{0,\gamma+1}, \quad \Theta = 1, 2, \cdots, \qquad (13.93)$$

并且对于刚性方程，h 由 (13.73) 来选取．而对于非刚性方程，h 由 (13.74) 来选取．

$g(t)$ 的另外一种近似方式是构造线性微分方程组

$$\frac{du(t)}{dt} = pu(t) + c, \ u(0) = u_0, \ u(t) \in R^{m_1}$$

使有

$$g(t) = Du(t) + b + r(g(t)),$$

其中 $r(g(t))$ 是近似的余项；D 是 $m \times m_1$ 矩阵．略去 $r(g(t))$，得到线性微分方程组

$$\frac{dx(t)}{dt} = Ax(t) + Du(t) + b, \ x(0) = x_0, \qquad (13.94)$$

$$\frac{du(t)}{dt} = pu(t) + c, \ u(0) = u_0.$$

对这个方程组，可以应用方法 (13.72)．另外，(13.94) 的系数矩阵是拟三角形的．在实际计算时，利用这一点可节省计算量及存贮量．

现在考虑变系数矩阵的微分方程组

$$\frac{dx}{dt} = A(t)x + g(t), \ x(t_0) = x_0. \qquad (13.95)$$

假定对于充分小的 ε_n，矩阵 $A(t)$ 满足

$$\|A(t) - A_n\| \leqslant \varepsilon_n, \ t_n \leqslant t \leqslant t_n + H_n, \ A_n = \text{常矩阵}.$$

与以前一样，假定在每个区间 $[t_n, t_n + H_n]$ 上，向量函数 $g(t)$ 能用幂多项式很好近似．有差分方程

$$z_{n+1} = \exp(A_n H_n) z_n + \sum_{\gamma=0}^{\nu} \int_0^{H_n} \exp(A_n H_n - A_n \tau) \frac{\tau^{\gamma}}{\gamma!} d\tau M_{\gamma}(t_n, H_n). \qquad (13.96)$$

它的矩阵系数每一步是变化的．在这种情形，若用前面考虑的算

法来确定这些系数矩阵，需要非常多的运算量．所以下面用另外一种方式来确定差分方程 (13.96) 中的矩阵系数．为了近似矩阵指数，应用下面的构造

$$\exp(A_nH_n) \simeq P_0(A_n, H_n)$$
$$= (E - \alpha A_nH_n)^{-l}(E - \beta_1 A_nH_n + \cdots$$
$$+ (-1)^{l-1}\beta_{l-1}A_n^{l-1}H_n^{l-1}), \quad l = 1, 2, \cdots, \tag{13.97}$$

矩阵指数的 $\gamma+1$ 重积分用矩阵函数 $P_{\gamma+1}(A_n, H_n)$ 代替，与 (13.83) 等价，根据 (13.97)，有递推式

$$P_\gamma(A_n, H_n) = A_nP_{\gamma+1}(A_n, H_n) + \frac{H_n^\gamma}{\gamma!}E, \tag{13.98}$$

$$\gamma = 0, \cdots, \nu+1.$$

利用这些代换后，差分方程 (13.96) 换成近似差分方程

$$\bar{z}_{n+1} = P_0(A_n, H_n)\bar{z}_n + \sum_{\gamma=0}^{\nu} P_{\gamma+1}(A_n, H_n)M_\gamma(t_n, H_n). \tag{13.99}$$

为了选取 (13.97) 右边的自由参数 α 和 β_i，首先要求它满足 $\nu+1$ 个条件

$$\lim_{t\to 0}\frac{\sum\limits_{k=0}^{s}\frac{t^k}{k!} - P_0(t)}{t^{s+1}} = \frac{1}{(s+1)!}, \quad s = 0, 1, \cdots, \nu < l, \tag{13.100}$$

其中 $p_0(t) = \dfrac{1 - \beta_1 t + \cdots + (-1)^{l-1}\beta_{l-1}t^{l-1}}{(1-\alpha t)^l}$. 条件 (13.100) 表示有理分式函数 $P_0(t)$ 的幂级数展开与指数 $\exp(t)$ 的 $\nu+1$ 项 Taylor 公式是一致的．除条件 (13.100) 外，余下的自由参数可选成极小化问题

$$\min_{(\alpha,\beta_i)}\max_{t\in[0,\infty)}|\exp(-t) - P_0(-t)|, \quad i = 1, 2, \cdots, l-1 \tag{13.101}$$

的解．

例如，取 $l=2$，$\nu=0$，根据 (13.97)、(13.98) 和 (13.100)，

我们有单参数构造（$s=0$）

$$\exp(A_nH_n)\simeq P_0(A_n,H_n)$$
$$=(E-\alpha A_nH_n)^{-2}[E-(2\alpha-1)A_nH_n]\tag{13.102}$$

和

$$\int_0^{H_n}\exp(A_n\tau)d\tau\simeq P_1(A_n,H_n)$$
$$=H_n(E-\alpha A_nH_n)^{-2}(E-\alpha^2A_nH_n).\tag{13.103}$$

将 (13.102) 和 (13.103) 代入方程 (13.99)，有

$$(E-\alpha A_nH_n)^2\bar{z}_{n+1}$$
$$=(E-(2\alpha-1)H_nA_n)\bar{z}_n+H_n(E-\alpha^2A_nH_n)M_0(t_n,H_n).\tag{13.104}$$

参数 α 可以由解问题 (13.101)

$$\min_{\alpha}\max_{t\in[0,\infty)}\left|\exp(-t)-\frac{1+(2\alpha-1)t}{1+2\alpha t+\alpha^2t^2}\right|\tag{13.105}$$

得到．由 (13.105) 得到的 $\alpha\simeq0.394$，而误差不超过 0.047（$t>0$）．

当 $\alpha=\frac{1}{3}\simeq0.333$ 时，差分方程 (13.104) 等价于下面的格式

$$\begin{aligned}z_{n+2/3}&=z_n+\frac{H_n}{3}[f(z_{n+2/3})+f(z_n)],\\ z_{n+1/3}&=z_{n+2/3}+\frac{H_n}{3}[f(z_{n+4/3})+f(z_{n+2/3})],\\ z_{n+1}&=\frac{1}{2}(z_{n+2/3}+z_{n+4/3}),\end{aligned}\tag{13.106}$$

$$f(z)=A_nz+M_0(t_n,H_n).$$

这时 $P_0(t)$ 近似 $\exp(t)$ 的误差是 0.126． 当消去 (13.106) 中的中间量 $z_{n+2/3}$ 和 $z_{n+4/3}$，格式 (13.106) 就变成 (13.104) 的形式．

应用 Rosenbrock 的二阶方法恰好对应于 $\alpha=0.293$ 的方程 (13.104)．这时近似的最大误差为 0.207．

§ 4.2 矩阵分解方法

考虑初值问题

$$\frac{dx}{dt}=f(t,x),\ f(t,x)\in C_{tx}^{(d,d)}(\Gamma),\ x(t)\in R^m,\ x(t_0)=x_0 \tag{13.107}$$

假定向量函数 $f(t,x)$ 定义的区域Γ对x是凸的. 引进变量t的离散值

$$t_n=t_0+\sum_{k=0}^{n-1}H_k,\ H_k>0, \tag{13.108}$$

这里n是整数. 由 Newton-Leibnitz 公式,可得

$$\begin{aligned}x(t_{n+1})&=x(t_n)+\int_0^{H_n}\frac{dx(t_n+\tau)}{d\tau}d\tau\\&=x(t_n)+\int_0^{H_n}\frac{dx(\rho)}{d\rho}\Big|_{\rho=t_{n+1}-\tau}d\tau,\end{aligned} \tag{13.109}$$

引进充分光滑的非奇矩阵 $\varphi_n(\tau)$ $(0\leqslant\tau\leqslant H_n)$, 这里足标$n$表示每一个离散步 $\varphi_n(\tau)$ 可以是不一样的. 将(13.109)记成形式

$$x(t_{n+1})-x(t_n)-\int_0^{H_n}\varphi_n^{-1}(\tau)\varphi_n(\tau)\frac{dx(\rho)}{d\rho}\Big|_{\rho=t_{n+1}-\tau}d\tau=0. \tag{13.110}$$

选取矩阵 $\varphi_n(\tau)$ 使得 $\varphi_n^{-1}(\tau)$容易积分,并且在t_n的邻域中,乘积

$$\varphi_n(\tau)\frac{dx(\rho)}{d\rho}\Big|_{\rho=t_{n+1}-\tau} \tag{13.111}$$

与τ的有限次多项式相差很小. 对(13.110)进行分部积分,并作一些运算后,得到

$$\begin{aligned}&x(t_{n+1})-x(t_n)-\left[\int_0^{H_n}\varphi_n^{-1}(\eta)d\eta+C_n\right]\varphi_n(H_n)\\&\quad\times\frac{dx(t_n)}{dt}+C_n\varphi_n(0)\frac{dx(t_{n+1})}{dt}\\&=\int_0^{H_n}\left[\int_0^{\tau}\varphi_n^{-1}(\eta)d\eta+C_n\right]\left[\varphi_n(\tau)\frac{d^2x(\rho)}{d\rho^2}\right.\\&\quad\left.-\frac{d\varphi_n(\tau)}{d\tau}\frac{dx(\rho)}{d\rho}\right]\Big|_{\rho=t_{n+1}-\tau}d\tau,\end{aligned} \tag{13.112}$$

在公式(13.112)中,矩阵 C_n 是任意的,它不依赖于 η. 矩阵 $\varphi_n(\tau)$ 和 C_n 的各种不同的选取，将得到各种不同的公式. 例如，取 $\varphi_n(\tau)=E$，等式(13.112)有形式

$$x(t_{n+1})-x(t_n)-(H_nE+C_n)\frac{dx(t_n)}{dt}+C_n\frac{dx(t_{n+1})}{dt}$$
$$=\int_0^{H_n}(E\tau+C_n)\frac{d^2x(\rho)}{d\rho^2}\Big|_{\rho=t_{n+1}-\tau}d\tau. \tag{13.113}$$

将(13.113)的右边略去，当 $C_n=0$ 时,得到 Euler 方法，当 $C_n=-H_nE$ 时,得到向后 Euler 方法,而当 $C_n=-\frac{H_n}{2}E$ 时,得到梯形公式.

对公式(13.112)继续进行分部积分，我们可以得到非常复杂的广义 Taylor 公式,称这个公式为矩阵分解式.

取 $\varphi_n(\tau)$ 为 $\exp(A_n\tau)$，则公式(13.112)取下面的等式

$$x(t_{n+1})-x(t_n)-\left[\int_0^{H_n}\exp(A_nH_n-A_n\tau)d\tau\right.$$
$$\left.+C_n\exp(A_nH_n)\right]\frac{dx(t_n)}{dt}+C_n\frac{dx(t_{n+1})}{dt}$$
$$=\int_0^{H_n}\left[\int_0^{\tau}\exp(A_n\tau-A_n\eta)d\eta+C_n\exp(A_n\tau)\right]$$
$$\times\left[\frac{d^2x(\rho)}{d\rho^2}-A_n\frac{dx(\rho)}{d\rho}\right]\Big|_{\rho=t_{n+1}-\tau}d\tau. \tag{13.114}$$

设 $C_n=0$，则恒等式(13.114)可简化,并改写成形式

$$x(t_{n+1})-x(t_n)-\int_0^{H_n}\exp(A_n\tau)d\tau\frac{dx(t_n)}{dt}$$
$$=\int_0^{H_n}\int_0^{\tau}\exp(A_n\eta)d\eta\left[\frac{d^2x(\rho)}{d\rho^2}-A_n\frac{dx(\rho)}{d\rho}\right]\Big|_{\rho=t_{n+1}-\tau}d\tau, \tag{13.115}$$

略去(13.115)的右边部分,我们得到数值积分的显式方法

$$z_{n+1}-z_n-\int_0^{H_n}\exp(A_n\tau)d\tau\, f(t_n,z_n)=0,\ z_n=z(t_n). \tag{13.116}$$

现在证明，如果方法(13.116)中的矩阵系数是计算精确的，则方法(13.116)将给出方程

$$\frac{dx(t_n+\tau)}{d\tau}=A_nx(t_n+\tau)+b_n,\ 0\leqslant\tau\leqslant H_n,\tag{13.117}$$

当 $\tau=H_n$ 时的精确解。为此，将(13.117)的右边部分代入(13.116)。根据(13.66)，我们有

$$\begin{aligned}&z_{n+1}-z_n-\int_0^{H_n}\exp(A_n\tau)d\tau(A_nz_n+b_n)\\&\quad=z_{n+1}-\exp(A_nH_n)z_n-\int_0^{H_n}\exp(A_n\tau)d\tau b_n=0,\end{aligned}\tag{13.118}$$

比较(13.117)和(13.118)，知它们是重合的。

当 $A_n=0$ 时，方法(13.116)为 Euler 方法，它对于常向量的积分是精确的。这样，在方法(13.116)中选取矩阵 A_n 可提高方法的精度或增大步长 H_n。下面考虑 A_n 的选取问题。设(13.107)是自守系统

$$\frac{dx}{dt}=f(x),\ x(t_0)=x_0\tag{13.119}$$

考虑它的变形

$$\frac{d^2x(t)}{dt^2}=\frac{\partial f(x)}{\partial x}\frac{dx(t)}{dt},\tag{13.120}$$

(13.119)可看成在解 $x(t)=x(t,t_0,x_0)$ 上的 $dx(t)/dt$ 的方程，并利用 Cauchy 矩阵将 dx/dt 写成

$$\frac{dx}{dt}=K(t,t_0)\frac{dx(t_0)}{dt},\tag{13.121}$$

其中 $K(t,t_0)$，满足微分方程

$$\frac{\partial K(t,t_0)}{\partial t}=\frac{\partial f(x)}{\partial x}K(t,t_0),\ K(t_0,t_0)=E.\tag{13.122}$$

利用 Cauchy 矩阵的性质，有

$$\frac{dx(t_n+\tau)}{d\tau}=K(t_n+\tau,t_n)\frac{dx(t_n)}{dt}.\tag{13.123}$$

积分(13.123)，得到

$$x(t_n+H_n)=x(t_n)+\int_0^{H_n}K(t_{n+\tau},t_n)d\tau\,\frac{dx(t_n)}{dt}. \quad (13.124)$$

因此，如果选取离散步长 H_n 有估计

$$\left\|\int_0^{H_n}K(t_{n+\tau},t_n)d\tau-\int_0^{H_n}\exp(A_n\tau)d\tau\right\|\leqslant\varepsilon, \quad (13.125)$$

其中 ε 为充分小的数，$\|\cdot\|$为所采用的矩阵模，则用方法 (13.116) 求解 (13.119) 时将有高的精度. 由此推得矩阵 A_n 可以取成为在离散步内某个 τ 值上的 Jacobi 矩阵. 通常取 $\tau=0$.

矩阵 A_n 取成为微分方程组的 Tacobi 矩阵的方法 (13.116) 称作系统拆线法. 现在建立用方法 (13.116) 求解 (13.107) 得到的近似解的误差方程. 假定近似解在 (13.107) 的精确解的存在区域中 ($\bar{G}\subset G$). 记

$$\varepsilon_n=\varepsilon(t_n)=x(t_n)-z(t_n)=x_n-z_n$$

应用等式

$$f(t_n,x_n)-f(t_n,z_n)=\int_0^1\frac{\partial f(t_n,x)}{\partial x}\bigg|_{x=z_n+\rho(x_n-z_n)}d\rho(x_n-z_n) \quad (13.126)$$

由 (13.115) 减去 (13.116)，对误差 ε_n 有差分方程

$$\varepsilon_{n+1}=\left[E+\int_0^{H_n}\exp(A_n\tau)d\tau\int_0^1\frac{\partial f(t_n,z_n+\rho\varepsilon_1)}{\partial z_n}d\rho\right]\varepsilon_n$$
$$+\int_0^{H_n}\int_0^{\tau}\exp(A_n\eta)d\eta\left[\frac{d^2x(\rho)}{d\rho^2}-A_n\frac{dx(\rho)}{d\rho}\right]_{\rho=t_{n+1}-\tau}d\tau. \quad (13.127)$$

应用 (13.66)，它可以变换成形式

$$\varepsilon_{n+1}=\left\{\exp(A_nH_n)+\int_0^{H_n}\exp(A_n\tau)d\tau\left[\int_0^1\frac{\partial f(t_n,z_n+\rho\varepsilon_n)}{\partial z_n}d\rho\right.\right.$$
$$\left.\left.-A_n\right]\right\}\varepsilon_n+\int_0^{H_n}\int_0^{\tau}\exp(A_n\eta)d\eta\left[\frac{d^2x(\rho)}{d\rho^2}\right.$$
$$\left.-A_n\frac{dx(\rho)}{d\rho}\right]_{\rho=t_{n+1}-\tau}d\tau. \quad (13.128)$$

当 $A_n=0$ 时，方程 (13.127) 和 (13.128) 变成 Euler 方法

的误差向量的差分方程

$$\varepsilon_{n+1}^{E}=\left(E+H_n\int_0^1\frac{\partial f(t_n,z_n+\rho\varepsilon_n)}{\partial z_n}d\rho\right)\varepsilon_n^{E}$$

$$+\int_0^{H_n}\tau\frac{d^2x(\rho)}{d\rho^2}\Bigg|_{\rho=t_{n+1}-\tau}d\tau, \tag{13.129}$$

其中 ε_n^E 表示用 Euler 方法的误差向量。

设向量模取为 $\|\varepsilon_n\|=\sup\limits_i|\varepsilon_n^{(i)}|$. 并取与这个模一致的矩阵模。对于具有实元素 a_{iin} 的矩阵 A_n，计算对数模

$$\|A_n\|=R_n=\sup_i\left(a_{iin}+\sum_{\substack{k=1\\i\neq k}}^{m}|a_{ikn}|\right) \tag{13.130}$$

利用对数模，可得到估计

$$\begin{aligned}&\|\exp(A_nH_n)\|\leqslant\exp(R_nH_n),\\&\left\|\int_0^{H_n}\exp(A_n\tau)d\tau\right\|\leqslant\int_0^{H_n}\exp(R_n\tau)d\tau,\\&\left\|\int_0^{H_n}\int_0^{\tau}\exp(A_n\eta)d\eta d\tau\right\|\leqslant\int_0^{H_n}\int_0^{\tau}\exp(R_n\eta)d\eta d\tau.\end{aligned} \tag{13.131}$$

对于 (13.107) 的特解引入数 α_n, β_n, γ_n, δ_n 来估计矩阵和向量的模，

$$\begin{aligned}&\left\|\int_0^1\frac{\partial f(t_n,z_n+\rho\varepsilon_n)}{\partial z_n}d\rho-A_n\right\|\leqslant\alpha_n,\\&\left\|\frac{\partial f(t_n+\tau,x(t_n+\tau))}{\partial x}-A_n\right\|\leqslant\beta_n,\end{aligned} \tag{13.132}$$

$$\|f(t_n+\tau,x(t_n+\tau))\|\leqslant\gamma_n,\left\|\frac{\partial f(t_n+\tau,x(t_n+\tau))}{\partial t}\right\|\leqslant\delta_n.$$

考虑

$$\frac{d^2x(t)}{dt^2}=\frac{\partial f(t,x)}{\partial x}\frac{dx}{dt}+\frac{\partial f(t,x)}{\partial t}, \tag{13.133}$$

得到估计

$$\|\varepsilon_{n+1}\|\leqslant\left[\exp(R_nH_n)+\alpha_n\int_0^{H_n}\exp(R_n\tau)d\tau\right]\|\varepsilon_n\|$$

$$+ (\beta_n \gamma_n + \delta_n) \int_0^{H_n} \int_0^{\tau} \exp(R_n \eta) d\eta d\tau. \tag{13.134}$$

现在考虑一些特殊情形．对于 $R_n = 0$，我们有估计

$$\|\varepsilon_{n+1}\| \leqslant (1 + \alpha_n H_n) \|\varepsilon_n\| + \frac{H_n^2}{2} (\beta_n \gamma_n + \delta_n). \tag{13.135}$$

在 $0 \leqslant -R_n \leqslant \alpha$ 的情形，也可以用公式 (13.135) 来估计，只要认为 $R_n = 0$．对于 $-R_n \geqslant \alpha_n$，成立公式

$$\exp(R_n H_n) + (\exp(R_n H_n) - 1) \frac{\alpha_n}{R_n} \leqslant 1. \tag{13.136}$$

因此，我们有估计式

$$\|\varepsilon_{n+1}\| \leqslant \|\varepsilon_n\| + \frac{\exp(R_n H_n) - 1 - R_n H_n}{R_n^2} (\beta_n \gamma_n + \delta_n), \tag{13.137}$$

对于 $R_n \geqslant \alpha_n$ 或者 $\alpha_n \geqslant R_n \geqslant 0$ 的情形，应用估计式 (13.134) 是合理的．

在上面的估计中，我们没有考虑到计算矩阵参数

$$\int_0^{H_n} \exp(A_n \tau) d\tau$$

的误差．但是下面的例子说明，即使是这种十分粗糙的估计也可以看出系统方法 (13.116) 求解刚性方程时相对于 Euler 方法的优越性．

例 13.2．考虑刚性方程

$$\frac{dx^{(1)}}{dt} = -501 x^{(1)} + 500 x^{(2)},$$

$$\frac{dx^{(2)}}{dt} = 500 x^{(1)} - 501 x^{(2)},$$

应用系统方法 (13.116) 求解．不取它的 Jacobi 矩阵作为 A_n，而取它的粗糙的近似值

$$\begin{pmatrix} -505 & 500 \\ 500 & -505 \end{pmatrix}$$ 其特征值为 -1001, -1

这时，我们有 $R_n = R = -5$，$\beta = \alpha = 4$，$\gamma = 1$，则利用误差

估计公式 (13.137)，进行一个步长的误差有

$$\|\varepsilon_1\| \leqslant \|\varepsilon_0\| + \frac{\exp(-5H_0) - 1 + 5H_0}{25}.$$

当 $H_0 = 0.1$，$\|\varepsilon_0\| = 0.01$ 时，$\|\varepsilon_1\| < 0.03$.

现在来建立隐式系统折线方法. 在公式 (13.114) 中选取

$$C_n = -\int_0^{H_n} \exp(-A_n\tau)d\tau. \tag{13.138}$$

在 C_n 的这种取法下，略去 (13.114) 的右边部分，得到

$$z_{n+1} - z_n - \int_0^{H_n} \exp(-A_n\tau)d\tau\, f(t_{n+1}, z_{n+1}) = 0. \tag{13.139}$$

但是对于刚性系统，(13.139) 中的矩阵系数计算不准，因此我们将其变换成别种形式. 将 (13.139) 乘上 $\exp(A_nH_n)$，加上和减去 $z_{n+1} - z_n$，则我们得到

$$z_{n+1} - z_n - [\exp(A_nH_n) - E](z_{n+1} - z_n)$$
$$- \int_0^{H_n} \exp(A_nH_n - A_n\tau)d\tau f_{n+1} = 0.$$

利用公式 (13.66)，我们有

$$z_{n+1} - z_n - \int_0^{H_n} \exp(A_n\tau)d\tau\, \{f_{n+1} - A_n(z_{n+1} - z_n)\} = 0. \tag{13.140}$$

为了与显式方法 (13.116) 形式上的一致性，将公式 (13.140) 变换成形式

$$z_{n+1} - z_n - \int_0^{H_n} \exp(A_n\tau)d\tau\, f(t_n, z_n)$$
$$- \int_0^{H_n} \exp(A_n\tau)d\tau[f(t_{n+1}, z_{n+1}) - f(t_n, z_n)$$
$$- A_n(z_{n+1} - z_n)] = 0. \tag{13.141}$$

为了求解非线性方程 (13.141)，只要应用简单迭代方法. 这时初始近似可以用由方法 (13.116) 得到的向量. 方法 (13.116) 和 (13.141) 的误差分析表明，几乎处处 (13.116) 的误差与 (13.141) 的误差符号上是相反的，并且在很多情形与离散步长成比例，至少在边界层以外是这样的. 因此，一方面可以用二个步长的计算来

检验结果的精度，另一方面可以构造比方法 (13.116) 和 (13.141) 更为精确的方法。

事实上，设用步长 H_n 由方法 (13.116) 得到的向量为 $\bar{z}(t_n+H_n)$

$$\bar{z}(t_n+H_n)=z(t_n)+\int_0^{H_n}\exp(A_n\tau)d\tau\, f(t_n,\, z(t_n)). \tag{13.142}$$

这里 $z(t_n)$ 假定是精确值。

用 $\frac{H_n}{2}$ 为步长由方法 (13.116) 计算二步得到向量

$$\bar{\bar{z}}\left(t_n+\frac{H_n}{2}\right)=z(t_n)+f(\tau_n,\, z(t_n))\int_0^{H_n/2}\exp(A_n\tau)d\tau,$$

$$\begin{aligned}\bar{\bar{z}}(t_n+H_n)\\ &=\bar{\bar{z}}\left(t_n+\frac{H_n}{2}\right)+\int_0^{H_n/2}\exp(A_n\tau)d\tau.\\ &\times f\left(t_n+\frac{H_n}{2},\ \bar{\bar{z}}\left(t_n+\frac{H_n}{2}\right)\right)\end{aligned}$$

由此，我们得到

$$\begin{aligned}\bar{\bar{z}}(t_n+H_n)=z(t_n)+\int_0^{H_n/2}\exp(A_n\tau)d\tau\\ \times\left[f_n+f\left(t_n+\frac{H_n}{2},z_n+\int_0^{H_n/2}\exp(A_n\tau)d\tau\ f_n\right)\right].\end{aligned} \tag{13.143}$$

现在假定 (13.143) 中的误差是 (13.142) 中的误差的一半.作向量

$$\tilde{z}(t_n+H_n)=2\bar{\bar{z}}(t_n+H_n)-\bar{z}(t_n+H_n), \tag{13.144}$$

则向量 $\tilde{z}(t_n+H_n)$ 中的误差比 $\bar{\bar{z}}(t_n+H_n)$ 和 $\bar{z}(t_n+H_n)$ 中的均要小。将公式 (13.143) 和 (13.142) 代入 (13.144)，并考虑到等式

$$\int_0^{H_n}\exp(A_n\tau)d\tau=\left[E+\exp\left(A_n\frac{H_n}{2}\right)\right]\int_0^{H_n/2}\exp(A_n\tau)d\tau,$$

$$\exp\left(A_n \frac{H_n}{2}\right) = E + \int_0^{H_n/2} \exp(A_n\tau)d\tau.$$

得到新的差分格式

$$\tilde{z}(t_n + H_n) = \bar{z}(t_n) + \int_0^{H_n/2} \exp(A_n\tau)d\tau \Big[2f\Big(t_n + \frac{H_n}{2}, z(t_n)$$

$$+ \int_0^{H_n/2} \exp(A_n\tau)d\tau\, f(t_n, z_n)\Big)$$

$$- \int_0^{H_n/2} \exp(A_n\tau)d\tau\, A_n f(t_n, z_n)\Big]. \tag{13.145}$$

不难验证，方法(13.145)对于求解方程(13.117)是精确的．当 $A_n = 0$ 时，它退化成二阶 Runge-Kutta 方法。因此，方法(13.145)称作是二阶系统方法.

若取 A_n 为离散步长 H_n 区间上(13.107)的 Jacobi 矩阵的近似，利用 $\exp(A_nH_n)$ 及其积分可以引进系统方法类．当 $A_n = 0$ 时，它们退化成对应阶的经典方法．构造的系统方法应该满足下面的条件

1) 如果 $A_n = 0$，则 ν 阶系统方法对于积分 $(\nu - 1)$ 阶代数多项式是精确的.

2) 任意阶的系统方法对于求解方程(13.117)是精确的.

一类系统方法可以用下面的方式来构造．对恒等式分部积分 $\nu - 1$ 次，我们得到下面的矩阵分解式（广义 Taylor 公式）

$$x(t_{n+1}) - x(t_n) - \int_0^{H_n} \exp(A_n\tau)d\tau \frac{dx(t_n)}{dt}$$

$$- \int_0^{H_n}\int_0^{\tau_2} \exp(A_n\tau_1)d\tau_1 d\tau_2\left[\frac{d^2x(t_n)}{dt^2} - A_n \frac{dx(t_n)}{dt}\right]$$

$$\cdots\cdots$$

$$-\int_0^{H_n}\int_0^{\tau_\nu}\cdots\int_0^{\tau_2} \exp(A_n\tau_1)d\tau_1\cdots d\tau_\nu\left[\frac{d^\nu x(t_n)}{dt^\nu} - \frac{A_n d^{\nu-1}x(t_n)}{dt^{\nu-1}}\right]$$

$$= \int_0^{H_n}\int_0^{\tau}\int_0^{\tau_\nu}\cdots\int_0^{\tau_2} \exp(A_n\tau_1)d\tau_1\cdots d\tau_\nu\Big[\frac{d^{\nu+1}x(\rho)}{d\rho^{\nu+1}}$$

$$- A_n \frac{d^\nu x(\rho)}{d\rho^\nu}\Big]_{\rho=t_{n+1}-\tau} d\tau. \tag{13.146}$$

显然，如果令 $A_n=0$，由 (13.93) 得到 Taylor 公式。

略去 (13.146) 的右边部分，得到 ν 阶的系统方法。利用等式

$$\int_0^{H_n}\int_0^{\tau_k}\cdots\int_0^{\tau_2}\exp(A_n\tau_1)d\tau_1\cdots d\tau_k$$
$$=\int_0^{H_n}\exp(A_nH_n-A_n\tau)\frac{\tau^{k-1}}{(k-1)!}d\tau. \qquad (13.147)$$

可以将它写成更紧凑的形式

$$z(t_{n+1})-z(t_n)-\int_0^{H_n}\exp(A_nH_n-A_n\tau)d\tau\frac{dz(t_n)}{dt}$$
$$-\sum_{k=2}^{\nu}\int_0^{H_n}\exp(A_nH_n-A_n\tau)\frac{\tau^{k-1}}{(k-1)!}d\tau$$
$$\times\left[\frac{d^kz(t)}{dt^k}-A_n\frac{d^{k-1}z(t)}{dt^{k-1}}\right]_{t=t_n}=0. \qquad (13.148)$$

不难验证，若给定方程

$$\frac{dx(t_n+\tau)}{d\tau}=A_nx(t_n+\tau)+\sum_{k=0}^{\nu-1}\frac{\tau^k}{k!}g_{kn},$$
$$0\leqslant\tau\leqslant H_n, \qquad (13.149)$$

其中 A_n 和 g_{kn} 不依赖于 τ，则由矩阵 A_n 的指数，方法 (13.148) 将给出 (13.149) 的精确解。因而它是 ν 阶的系统方法。

按照系统方法的定义，它们至少对于 $\nu=1$ 的方程 (13.149) 是精确的，因此方程 (13.148) 具有更大的精确方程的范围。

当 $\nu=2$ 时，经常取下面形式的方法 (13.148)

$$z_{n+1}-z_n-\int_0^{H_n}\exp\left(\frac{\partial f_n}{\partial z}\tau\right)d\tau f_n-$$
$$\int_0^{H_n}\int_0^{\tau}\exp\left(\frac{\partial f_n}{\partial z_n}\eta\right)d\eta d\tau\frac{\partial f_n}{\partial t}=0 \qquad (13.150)$$
$$\frac{\partial f_n}{\partial z}=\frac{\partial f(t_n,z_n)}{\partial z},\quad\frac{\partial f_n}{\partial t}=\frac{\partial f(t_n,z_n)}{\partial t}.$$

(13.150) 中的矩阵系数仍用 §4.1 中的方法计算，而偏导数 $\partial f_n/\partial t$ 可以由近似式

$$H_n \frac{\partial f_n}{\partial t} \simeq f(t_{n+1}, z_n) - f(t_n, z_n)$$

来计算.

§5 线性多步平均算法

Dahlquist[48] 建议的以不同的步长用梯形方法进行积分，然后用 Richardson 整体外插，可得到 A 稳定的高阶数值积分方法. 这种方法采用了平滑过程后是有效的. 但它有一个缺点，只在基本点上进行外插，而在基本点以外的节点上的计算结果在外插过程中是没有用的. 这似乎是一种浪费. 事实上，在上述的计算过程中，积分方法是固定的，步长作为一个可变参数，而 Richardson 外插可以看成为对不同参数(步长)的计算结果的某种平均过程，从而得到高阶的计算结果. 为了克服上面提到的缺点，可以将积分步长固定，而在保持 A 稳定的条件下设法在方法中引进可变参数，然后以固定的步长而用不同的参数的方法进行积分，最后将积分结果进行某种平均来得到高阶的结果. 这就是 Liniger 和 Odeh[79] 1972 年提出的方法的基本思想. 本节主要讨论 Adams 型线性多步平均算法.

考虑求解常微分方程初值问题

$$y' = f(y), \quad y(0) = y_0. \tag{13.151}$$

设(13.151)的精确解 $y(t)$ 是在 $t \geqslant 0$ 上的无限连续可微函数. 考虑 Adams 型的线性 k 步方法

$$y_{n+1} - y_n = h\left[cf_{n+1} + (1-c)f_n + \sum_{j=1}^{k-1}(b_j^* - c)\nabla^j f_n\right], \tag{13.152}$$

其中 ∇^j 是第 j 阶的向后差分算子，$\nabla f_n = f_n - f_{n-1}$. 对(13.152)，引进一个算子

$$\Omega = \Delta - hD\left[c\Delta + I + \sum_{j=1}^{k-1}(b_j^* - c)\nabla^j\right], \tag{13.153}$$

其中 Δ 是一阶向前差分算子，$\Delta f_n = f_{n+1} - f_n$，$D = \frac{d}{dt}$.

现在用下面的方式来确定算子 Ω 的精度阶. 将函数 $y(t)$ 及其差分在 $t = t_n$ 处展开，并用记号"$\simeq$"表示具有误差 $O(h^k)$ 的相等. 因此有

$$\Delta \simeq \sum_{i=1}^{k} (i!)^{-1}(hD)^i, \tag{13.154}$$

$$\nabla^j \simeq \sum_{i=j}^{k} \gamma_{ij}(hD)^i, \tag{13.155}$$

其中 γ_{ij} 是形式乘法

$$\left(D - \frac{1}{2!} D^2 + \frac{1}{3} D^3 - \cdots\right)^j = D^j + \cdots,$$

得到的右端项的系数. 将 (13.154)、(13.155) 代入 (13.153)，我们得到

$$\Omega \simeq \sum_{i=2}^{k} \nu_i (hD)^i, \tag{13.156}$$

其中

$$\nu_i = (i!)^{-1} - \sum_{j=1}^{i-1} \gamma_{i-1,\, j} b_j^*, \quad 2 \leqslant i \leqslant k. \tag{13.157}$$

量 ν_i 与参数 b_j^* 之间的关系式 (13.157) 与参数 c 无关. 因此象 (13.152) 或 (13.153) 那样引进参数 c 将不影响精度阶，而只影响方法的稳定性.

在 (13.156) 中若有 $\nu_2 = \nu_3 = \cdots = \nu_p = 0$，则算子 Ω 具有精度阶 p. 若 $\nu_2 \neq 0$，则算子 Ω 的精度阶为 1. 相应地方法 (13.152) 具有阶 p 或 1.

对于所有 i 有 $\gamma_{ii} = 1$，再由 (13.157) 中 b_j^* 的系数矩阵是三角形矩阵，对任意给定的 ν_i 的组，可以唯一地解出 b_j^*. 特别对任意的 p，$2 \leqslant p \leqslant k$，我们总可以选取常数 b_j^*，使得有 $\nu_i = 0$，$i = 2, \cdots, p$，这时算子 Ω 将有精度阶 p. 下面令 b_j 为使 (13.157) 中的所有 ν_i 均为零的 b_j^* 的解. 这时方法的阶

$p=k$. 但是当 $p>2$ 时,这个方法不可能是 A 稳定的，因而用它来求解刚性方程将是不合适的.

现在我们考虑另外一种应用方法 (13.152) 的方式，使能同样得到 $p=k$ 阶方法 (13.152) 的效果. 将 (13.152) 中的常数 b_j^* 取成为

$$
\begin{aligned}
b_j^* &= b_j, & j &= 1,\cdots, p-1, \\
b_j^* &= b_j + u_{j-p+1}, & j &= p,\cdots, k-1,
\end{aligned} \tag{13.158}
$$

其中 $u_1,\cdots,u_d$, $d=k-p$ 是任意常数.容易看出,这时(13.156)可改写成

$$
Q \simeq (hD)^p \sum_{i=1}^{d} \mu_i (hD)^i, \tag{13.159}
$$

其中

$$
\mu_i = -\sum_{j=1}^{i} \gamma_{p-1+i,\, p-1+j} u_j, \quad i=1,\cdots,d. \tag{13.160}
$$

因此，对于 $u_1,\cdots,u_d$ 的任何值，Q 的阶 $\geqslant p$. 每一个这样的参数组组成一个 d 维向量 $u=(u_1,\cdots,u_d)^T$. 考虑 $m+1$ 个这样的向量 $u_\rho=(u_{1,\rho},\cdots,u_{d,\rho})^T$, $\rho=1,\cdots,m+1$. 并且对每个 ρ，定义算子 Q_ρ 为

$$
\begin{aligned}
Q_\rho = \Delta - hD\Bigg[c\Delta + I + \sum_{j=1}^{p-1}(b_j-c)\nabla^j \\
+ \sum_{j=p}^{k-1}(b_j - c + u_{j-p+1,\rho})\nabla^j \Bigg],
\end{aligned} \tag{13.161}
$$

再作算子

$$
Q' = \sum_{\rho=1}^{m+1} \nu_\rho Q_\rho. \tag{13.162}
$$

对于给定的向量 u_ρ, $\rho=1,\cdots,m+1$, 我们来说明如何选取权 ν_ρ, 使得 Q' 具有阶 $p+m$. 由 (13.159), (13.160), (13.162), 我们得到

$$
Q' \simeq (hD)^p \sum_{i=1}^{d} \pi_i (hD)^i, \tag{13.163}
$$

其中

$$\pi_i = \sum_{\rho=1}^{m+1} \nu_\rho \mu_{i,\rho}, \tag{13.164}$$

$\mu_{i,\rho}$ 是 (13.160) 中的 u_i 取值 $u_{i,\rho}$ 得到的值. 由此可以看到,为了 Ω' 有阶 $p+m$, 必须要有

$$\pi_i = 0, \quad i = 1, \cdots, m. \tag{13.165}$$

可以证明, (13.165) 与

$$\sum_{\rho=1}^{m+1} \nu_\rho u_{i,\rho} = 0, \quad i = 1, \cdots, m \tag{13.166}$$

是等价的. 对 ν_ρ 再加上正则化条件

$$\sum_{\rho=1}^{m+1} \nu_\rho = 1. \tag{13.167}$$

这样,如果 ν_ρ 满足 (13.166) 和 (13.167), 则算子 Ω' 和相应的方法 (13.152) 将具有阶 $p+m$.

容易看出 (13.166)、(13.167) 对 ν_ρ, $\rho = 1, \cdots, m+1$ 的唯一可解性等价于m个向量 $u_\rho - u_{m+1}$ 可展成m维线性空间. 对于$m=2$的特殊情形, 唯一可解性条件是以二维向量 u_1, u_2, u_3 作为顶点组成一个非退化三角形.

方法 (13.152) 当 $c \neq 0$ 时是隐式的, 每积分一步需要求解一个方法所产生的非线性方程组. 对于刚性方程, 在求解这个非线性方程组时, 往往需要采用 Newton-Raphson 迭代方法. 这就需要计算右函数 $f(y)$ 的 Jacobi 矩阵及其有关的逆矩阵. 在具体设计数值软件时,还可以将这个 Jacobi 矩阵直接加到数值积分公式中去,构造出更适合于求解的公式.

将 (13.152) 中出现的差分线性化, 并且将相应的 Jacobi 矩阵用 Jacobi 矩阵 $\partial f/\partial y$ 在 (t_n, y_n) 处的近似值 J_n 代替,得到下面的公式

$$y_{n+1} = y_n + chJ_n \left[\Delta y_n - \sum_{j=1}^{k-1} \nabla^j y_n \right]$$

$$+ h\left[f_n + \sum_{i=1}^{k-1} b_i^* \nabla^i f_n\right], \tag{13.168}$$

这种公式称作是线性隐式的. 若方法 (13.152) 的阶 $p \leqslant k$ 时, 公式 (13.152) 中以 c 为因子的项将不影响公式的精度阶. 同样公式 (13.168) 中含 c 的项也不影响公式的精度阶. 所以进行精度分析时, 公式 (13.152) 和公式 (13.168) 是等同的. 容易看出进行 A 稳定分析时, 这两个公式也是等同的. 公式 (13.168) 对 y_{n+1} 是线性的, 只要矩阵 $[I - chJ_n]$ 的逆存在, 立即可以求得

$$y_{n+1} = \left[I - chJ_n\right]^{-1}\left\{y_n - chJ_n \sum_{i=1}^{k-1} \nabla^i y_n\right.$$

$$\left. + h\left[f_n + \sum_{i=1}^{k-1} b_i^* \nabla^i f_n\right]\right\}. \tag{13.169}$$

公式 (13.168) 也可以象公式 (13.152) 一样进行平均.

上述算法在并行处理计算机上实现是很方便的. 按前面叙述的方式选取参数向量 u_ρ 和权系数 v_ρ, $\rho = 1, \cdots, m+1$. 对于每个 ρ, 由 (13.169) 求得 $y_{n+1,\rho}$, 再按

$$y_{n+1} = \sum_{\rho=1}^{m+1} v_\rho y_{n+1,\rho} \tag{13.170}$$

求出平均解 y_{n+1}. 计算的示意框图由图 13.1 表示.

下面举几个具体的算法的例子.

例 13.3 考虑 Adams 型算法

$$y_{n+1} = y_n + h\left[cf_{n-1} + (1-c)f_n + \left(\frac{1}{2} + r - c\right)\nabla f_n\right]. \tag{13.171}$$

它的精度阶 $p = 2$, $k = 2$. 将其写成

$$y_{n+1} = y_n + h\left[-\left(\frac{1}{2} - c + r\right)f_{n-1}\right.$$

$$\left. + \left(\frac{3}{2} - 2c + r\right)f_n + cf_{n+1}\right],$$

对应的生成多项式为

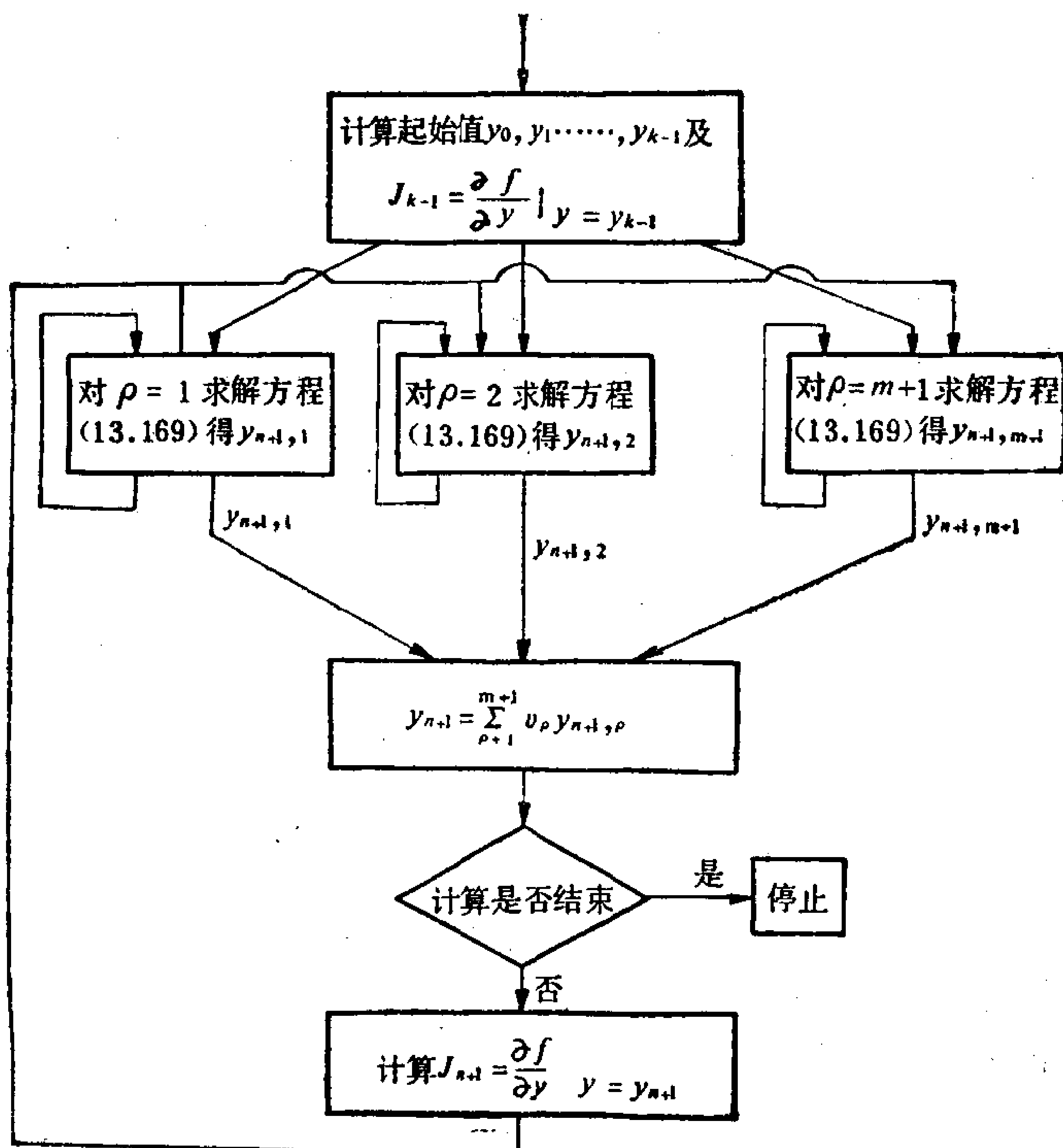


图 13.1　平均算法示意框图

$$\rho(\zeta)=\zeta^2-\zeta, \tag{13.172}$$

$$\sigma(\zeta)=c\zeta^2+\left(\frac{3}{2}-2c+r\right)\zeta-\left(\frac{1}{2}-c+r\right). \tag{13.173}$$

下面应用第二章的定理 2.3 来确定保证公式 (13.171) 为 A 稳定的参数 r 和 c 的区域．作变换

$$\zeta=\frac{z+1}{z-1}, \tag{13.174}$$

并令

$$S(z)=\left(\frac{z-1}{2}\right)^2\sigma(\zeta),$$

$$\sigma(\zeta)=\frac{1}{4}[z^2+(1+2r)z+(-2+4c-2r)].$$

(13.175)

为要求 $\sigma(\zeta)$ 的零点均在单位圆内部等价于要求多项式 $S(z)$ 的零点均在左半平面内. 由 Routh 准则,这又等价于

$$1+2r>0, \tag{13.176}$$

$$1-2c+r<0. \tag{13.177}$$

由于

$$\alpha_0=0,\quad \beta_0=-\left(\frac{1}{2}-c+r\right),$$

$$\alpha_1=-1,\quad \beta_1=\left(\frac{3}{2}-2c+r\right),$$

$$\alpha_2=1,\quad \beta_2=c.$$

可求得

$$\gamma_0=-\frac{3}{2}+3c-r,$$

$$\gamma_1=2-4c+2r,$$

$$\gamma_2=-\left(\frac{1}{2}-c+r\right).$$

于是对应于公式(13.171)的多项式 $P_2(\xi)$ 为

$$P_2(\xi)=(\xi-1)[(-1+2c-2r)\xi+(1-2c)], \tag{13.178}$$

因此要求条件(2.15)成立等价于要求

$$(-1+2c-2r)\xi+(1-2c)\leqslant 0,\quad -1\leqslant\xi\leqslant 1 \tag{13.179}$$

成立. 令 $\xi=1$ 和 $\xi=-1$, 即推得条件

$$-1+2c\geqslant r\geqslant 0, \tag{13.180}$$

满足(13.180)的点 (r,c) 在楔形 $A1$ 中(见图13.2).

容易验证,除边界的点外,楔形 $A1$ 中的点还满足条件(13.170)、(13.177). 因此,如果 $0\leqslant r<2c-1$, 则方法(13.171)是 A 稳定的.

例 13.4 考虑 Adams 型算法

$$y_{n+1}=y_n+h\Big[c\Delta f_n+f_n+\left(\frac{1}{2}-c\right)\nabla f_n$$

$$+\left(\frac{5}{12}-c+r\right)\nabla^2 f_n\Big], \tag{13.181}$$

这个方法的精度阶 $p=2$，$k=3$. 将其写成通常的形式为

$$y_{n+1}=y_n+h\Big[cf_{n+1}+\left(\frac{23}{12}-3c+r\right)f_n$$
$$-\left(\frac{8}{6}-3c+2r\right)f_{n-1}+\left(\frac{5}{12}-c+r\right)f_{n-2}\Big], \tag{13.182}$$

其对应的生成多项式为

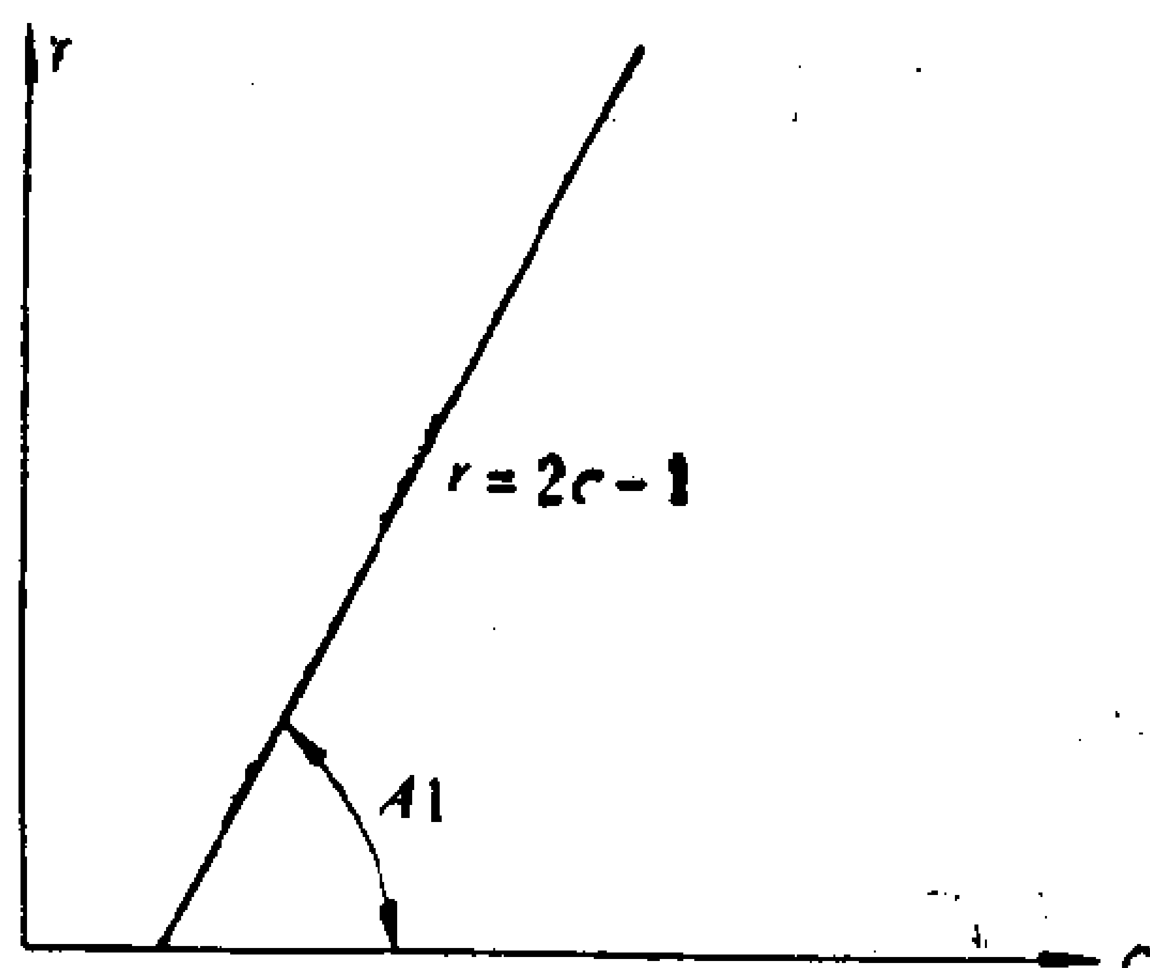


图 13.2 算法 (13.171) 的 A 稳定的参数区域

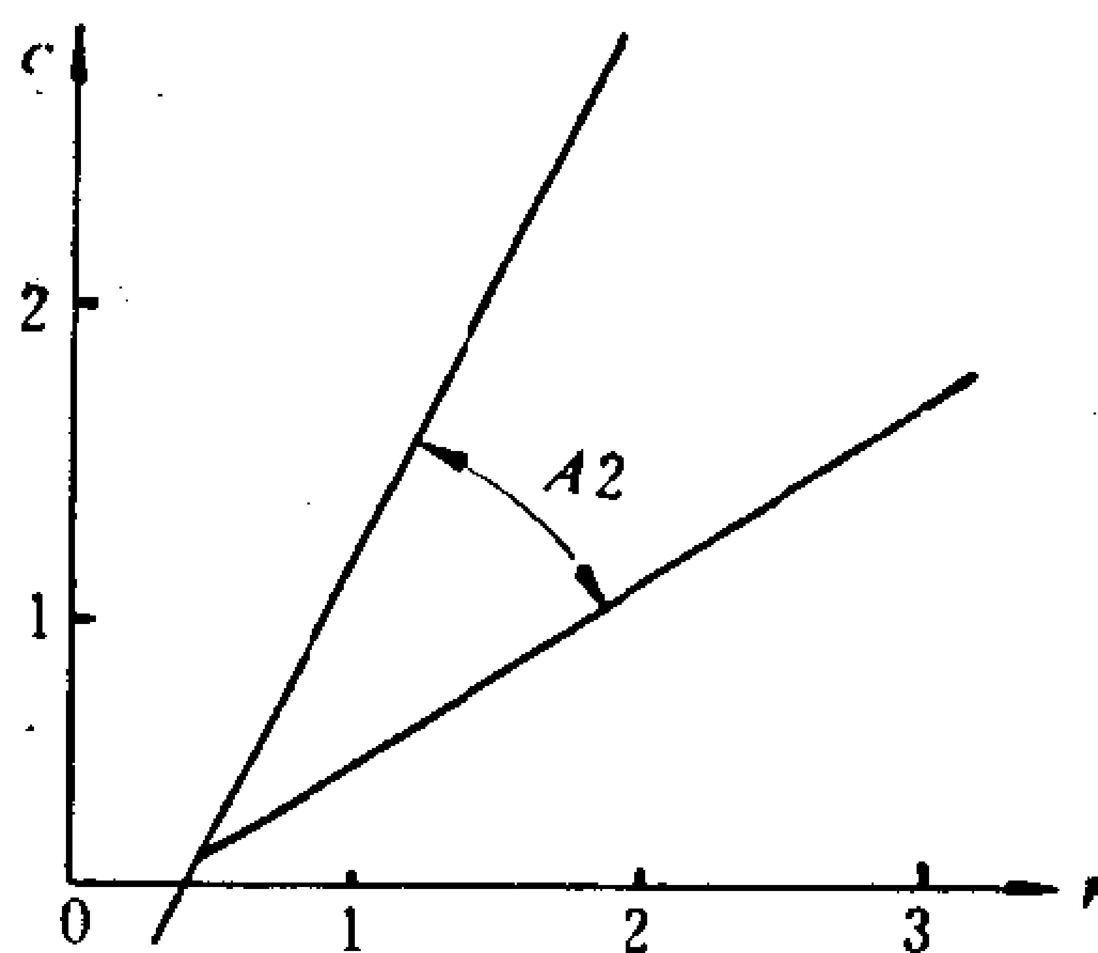


图 13.3 算法 (13.181) 的 A 稳定的参数区域

$$\rho(\zeta)=\zeta^3-\zeta^2,$$
$$\sigma(\zeta)=c\zeta^3+\left(\frac{23}{12}-3c+r\right)\zeta^2-\left(\frac{8}{6}-3c+2r\right)\zeta$$
$$+\left(\frac{5}{12}-c+r\right). \tag{13.183}$$

作变换 (13.174)，并令

$$S(z)=\left(\frac{z-1}{2}\right)^3\sigma(\zeta)$$
$$=\frac{1}{24}[3z^3+6z^2+(2+12r)z+(-11+24c-12r)]. \tag{13.184}$$

应用 Routh 准则，要求 $S(z)$ 的根均在左半平面等价于 r 满足不

等式

$$-\frac{5}{12}+\frac{2}{3}c<r<-\frac{11}{12}+2c. \tag{13.185}$$

对应于方法 (13.181) 的函数 $P_3(\xi)$ 为(由 (2.16) 定义)

$$P_3(\xi)=(\xi-1)^2\left[\left(\frac{5}{3}-4c+4r\right)\xi-\left(\frac{1}{6}-2r\right)\right], \tag{13.186}$$

因此要求对 $-1\leqslant\xi\leqslant 1$ 有 $P_3(\xi)\geqslant 0$ 等价于要求有

$$\left(\frac{5}{3}-4c+4r\right)\xi-\left(\frac{1}{6}-2r\right)\geqslant 0,\quad -1\leqslant\xi\leqslant 1.$$

取 $\xi=1$, $\xi=-1$ 立即推得不等式

$$2c-\frac{11}{12}\geqslant r\geqslant\frac{2}{3}c-\frac{1}{4}. \tag{13.187}$$

满足 (13.187) 的点 (r, c) 组成图 13.3 中的楔形 $A2$. 与 (13.185) 相比较，可知对于除去上界 $r=2c-\frac{11}{12}$ 后的楔形 $A2$ 中的点 (r, c), 方法 (13.181) 是 A 稳定的.

在 $A2$ 中选取适当的 r_ρ 及权系数 ν_ρ. 利用本节所述的平均过程可得到 $p=3$ 的计算结果.

例 13.5 考虑 Adams 型算法

$$y_{n+1}=y_n+h\left[c\Delta f_n+f_n+\left(\frac{1}{2}-c\right)\nabla f_n\right.$$
$$\left.+\left(\frac{5}{12}-c+r\right)\nabla^2 f_n+\left(\frac{3}{8}-c+s\right)\nabla^3 f_n\right], \tag{13.188}$$

将其写成通常的形式后，其对应的生成多项式为

$$\rho(\zeta)=\zeta^4-\zeta^3, \tag{13.189}$$

$$\sigma(\zeta)=c\zeta^4+\left(\frac{55}{24}-4c+r+s\right)\zeta^3-\left(\frac{59}{24}-6c+2r+3s\right)\zeta^2$$
$$+\left(\frac{37}{24}-4c+r+3s\right)\zeta-\left(\frac{3}{8}-c+s\right). \tag{13.190}$$

这个方法的精度阶为 $p=2$, $k=4$. 应用第二章的定理 2.3, 作变换 (13.174), 并令

$$S(z) = \left(\frac{z-1}{2}\right)^4 \sigma(\zeta)$$

$$= \frac{1}{48}[3z^4 + 9z^3 + (8 + 12r)z^2 + 24sz$$

$$+ (-20 + 48c - 12r - 24s)], \qquad (13.191)$$

应用 Routh 准则，要求 $S(z)$ 的根均在左半平面内等价于成立不等式

$$c_0 = 2 + 3r - 2s > 0, \qquad (13.192)$$

$$d_0 = (15 - 36c + 9r + 34s + 24rs - 16s^2)/(2 + 3r - 2s) > 0, \qquad (13.193)$$

$$e_0 = -5 + 12c - 3r - 6s > 0. \qquad (13.194)$$

对于任意的 $c \neq 0$ 和 $c_0 \neq 0$，$d_0 = 0$ 表示一非退化的双曲线，其中心在$\left(-\frac{23}{12}, -\frac{3}{8}\right)$处，渐近线为 $As1$: $s = -\frac{3}{8}$ 和 $As2$: $s = \frac{3}{2}r + \frac{5}{2}$. $As2$ 与 $c_0 = 0$ 是平行的. 这双曲线与 $c_0 = 0$ 和 $e_0 = 0$ 交在点 $p_2 = \left(-\frac{11}{12} + c, -\frac{3}{8} + \frac{3}{2}c\right)$，和 $e_0 = 0$ 的另一个交点为 $p_1 = \left(-\frac{5}{3} + 4c, 0\right)$. 当 $c = \frac{1}{4}$ 时，这两个交点将重合. 如果 $c > 0$，双曲线 $d_0 = 0$ 的另一支（位于 $s < -\frac{3}{8}$ 中）可以不考虑，因为它所构成的区域与 $c_0 > 0$ 不相交. 下面还将证明：如果 $c < \frac{1}{2}$，则条件 (2.15) 将不满足. 因此我们仅限于讨论 $c \geq \frac{1}{2}$ 的情形.

对应于方法 (13.188) 的多项式 $P_4(\xi)$ 为

$$P_4(\xi) = (\xi - 1)^2 Q(\xi), \qquad (13.195)$$

其中

$$Q(\xi) = (-9 + 24c - 24s)\xi^2 + (5 - 12c + 12r)\xi + (4 - 12c + 6r + 12s) \qquad (13.196)$$

因此，为要求当$-1 \leqslant \xi \leqslant 1$时$P_4(\xi) \geqslant 0$等价于要求当$-1 \leqslant \xi \leqslant 1$时有

$$Q(\xi) \geqslant 0. \tag{13.196'}$$

对于固定的$c > 0$和固定的ξ、$-1 \leqslant \xi \leqslant 1$，$Q(\xi) = 0$为参数$r$和$s$平面上的一条直线．当$\xi$在$-1 \leqslant \xi \leqslant 1$中变化时，这些直线的包络为椭圆（见图 13.4）．即

$$6r^2 + 24rs + 48s^2 + (14 - 36c)r + (34 - 96c)s + \frac{1}{24}(13 - 36c)^2 = 0, \tag{13.197}$$

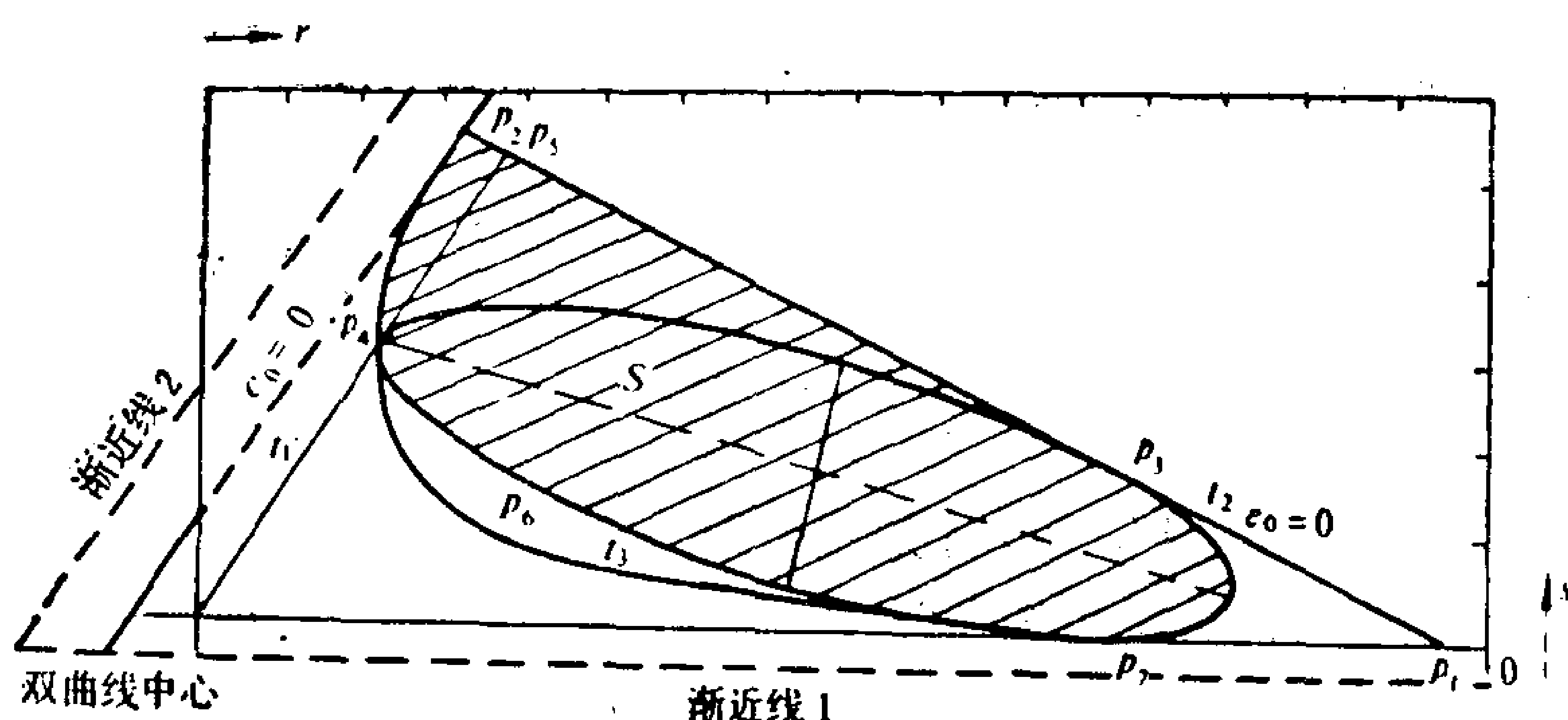


图 13.4　$c = 4$ 时算法（13.188）的A稳定的参数区域

它位于$p_4 = \left(-\frac{13}{60} + \frac{3}{5}c, \ -\frac{13}{40} + \frac{9}{10}c\right)$（它是对应于$\xi = +1$的直线$Q(1) = 0$与椭圆（13.197）的切点，这直线记成$t_1$）和

$$p_3 = \left(-\frac{17}{12} + 3c, -\frac{1}{8} + \frac{1}{2}c\right)$$

（它是对应于$\xi = -1$的直线$Q(-1) = 0$与椭圆（13.197）的切点，这直线记成t_2）之间的部分．t_2与$e_0 = 0$重合．通过计算$s = s(\xi)$（切点的s坐标）并且证明它在$P_4(\xi = 1)$和

$$P_7 = \left(-\frac{17}{12} + 3c, \frac{1}{8}\right)$$

（对应于$\xi = -\frac{1}{2}$的直线$Q\left(-\frac{1}{2}\right) = 0$与椭圆（13.197）的切

点)之间是减的，而在

$$P_7\left(\xi=-\frac{1}{2}\right) \text{ 和 } P_3(\xi=-1)$$

之间是增的. 所以导数 $ds/d\xi$ 在 $-1\leqslant\xi\leqslant1$ 中存在零点.

由(13.196)，和 $Q(0)\geqslant 0$ 解得

$$s\geqslant-\frac{1}{2}r+c-\frac{1}{3}, \tag{13.198}$$

而由 $Q(-1)\geqslant 0$ 解得

$$s\leqslant\left(-\frac{5}{6}+2c\right)-\frac{1}{2}r. \tag{13.199}$$

为保证(13.198)和(13.199)是相容的，必须有

$$c-\frac{1}{3}\leqslant-\frac{5}{6}+2c,$$

因此仅当

$$c\geqslant\frac{1}{2}$$

时，(13.198)、(13.199)才是相容的.

由上面的分析可知在以 P_3 和 P_4 间的椭圆的下面部分，直线 t_1 和直线 t_2 所界定的闭集 $\bar{S}$ 上，$Q(\xi)\geqslant 0$，因而条件(2.15)将满足.

现在我们证明在将 $\bar{S}$ 在直线 t_2 上的闭边界部分 $[p_3, p_5]$ 去掉后的开闭集合 S 上，算法(13.188)是 A 稳定的. 这只要证明在 S 上不等式(13.192)—(13.194)均满足. 首先不等式 $Q(-1)>0$ 与(13.194)定义同样的开半平面(切线 t_2 的下半部分)，包含所有的 S.

其次，S 的所有点在以 t_1 为界的半平面 $s\leqslant\frac{3}{2}r$ 中，因此满足较弱的不等式(13.192). 第三，为了看出 S 的所有点在由不等式 $d_0>0$ 所确定的区域中，只要证明(如果 p_3 在 $d_0>0$ 中) S 的边界与双曲线不相交. 由于

$$p_5=\left(-\frac{5}{12}+c,\ -\frac{5}{8}+\frac{3}{2}c\right)$$

是 t_1 和 t_2 的交点，容易证明区间 $[p_3, p_5]$ 严格地位于 t_2 与双曲线的二个交点 p_1 和 p_2 之间. 因此 $[p_3, p_5]$ 全部在 $d_0 > 0$ 所定的区域中. 类似地，p_4 和 p_5 均在 t_1 与双曲线的单个交点的 $(d_0 > 0)$ 一边，因而整个区间 $[p_4, p_5]$ 也在 $d_0 > 0$ 所确定的区域中.

最后，为了说明 S 的边界的椭圆部分与双曲线不交，我们证明这对整个椭圆均是成立的. 由 (13.197) 和 $d_0 = 0$ 消去 r，并令 $s = 3\sigma/4$，则椭圆和双曲线的交点一定对应于 σ 的方程

$$(72\sigma^2 - 24\sigma + 2)c^2 + (-120\sigma^3 - 28\sigma^2 + 14\sigma - 1)c$$
$$+ \left(50\sigma^4 + 40\sigma^3 + 3\sigma^2 - 2\sigma + \frac{1}{8}\right) = 0 \qquad (13.200)$$

的实根. 但是 (13.200) 对 c 的判别式等于 $-21\sigma^2$. 因此，除非 $\sigma = 0$，对任何实 σ，任何 c 的实值均不满足 (13.200)，这等价于对任何实值 c，方程 (13.200) 对 σ 无实解，除非 $\sigma = 0$. 但 $\sigma = 0$，对应于 $c = \frac{1}{4}$，我们不考虑这个 c 的值. 公式 (13.188) 含有三个参数 c, r, s，可以选取三组参数，使平均解达到的精度阶为 4. 选取参数通常要考虑到这样两点: (i) 保证对 ν_ρ 的唯一可解性; (ii) 得到的算法都是 A 稳定的. 除这两点外，还可以使解在 $q = \infty$ 远处具有最大可能的衰减，或在某种意义下使局部截断误差达到极小. 对于算法 (13.188) 可以取

$$c = 4,$$
$$(r_1, s_1) = (7, 2),$$
$$(r_2, s_2) = (5, 2),$$
$$(r_3, s_3) = (7, 1),$$

§6 块 方 法

我们知道在使用隐式 Runge-Kutta 方法时，是为了由 t_{n-1} 计算 t_n 处的方程的数值解，它同时计算了在 t_{n-1} 和 t_n 之间的一些

中间点上的值．当然在这些点上的值精度比在 t_n 处的精度差得多，但提高了 A 稳定公式的阶．将这种思想推广，构造相应的计算方法，使得这些中间点为我们所需要的节点．而在这些点上解的近似值也具有所需要的精度．这种方法与传统的方法不同，每计算一次得到的不是一个新的节点上的值，而是一组节点上的值．块方法就是具有这种性质的方法．当然，这种方法的构造思想与 Runge-Kutta 方法是不同的．下面我们着重介绍块隐式单步方法．

对于步长 $h\in(0, h_0]$，令 $t_k = t_0 + kh$，则按下述方式构造初值问题

$$y' = f(t, y),\ y(t_0) = y_0 \tag{13.201}$$

的解 $y(t)$ 在节点序列 $\{t_k\}$ 上的近似 $\{y_k\}$ 的方法称 r 块单步方法．在得到值 y_n，$n = mr$ 后，往后计算一步可同时得到 r 个值 $y_{n+1}, y_{n+2}, \cdots, y_{n+r}$．本节我们考虑由公式

$$\sum_{j=1}^{r} a_{ij}y_{n+j} = e_i y_n + hd_i f_n + h\sum_{j=1}^{r} b_{ij}f_{n+j},\ i = 1, \cdots, r \tag{13.202}$$

来构造 $y_{n+1}, \cdots, y_{n+r}$ 的块单步方法，其中 a_{ij}，e_i，d_i，b_{ij} 均是常数，$f_{n+i} = f(t_{n+i}, y_{n+i})$．

应用向量和矩阵的记号，记

$$A = (a_{ij}),\ B = (b_{ij}),\ e = (e_1, \cdots, e_r)^T,\ d = (d_1, \cdots, d_r)^T$$

和

$$Y_m = (y_{n+1}, \cdots, y_{n+r})^T,\ F(Y_m) = (f_{n+1}, \cdots, f_{n+r})^T,$$

则方法 (13.202) 可表成

$$AY_m = hBF(Y_m) + ey_n + hdf_n. \tag{13.203}$$

我们假定 $A = I$，而只讨论形状为

$$Y_m = hBF(Y_m) + ey_n + hdf_n \tag{13.204}$$

的方法，因为必要时还可以在 (13.202) 的两边乘上 A^{-1}，即可化成 (13.204) 的形状．

例 13.6 （Clippinger-Dimsdale 方法）

$$y_{n+1} - \frac{1}{2}y_{n+2} = \frac{1}{2}y_n + \frac{h}{4}f_n - \frac{h}{4}f_{n+2},$$

$$y_{n+2} = y_n + \frac{h}{3} f_n + \frac{4}{3} h f_{n+1} + \frac{h}{3} f_{n+2}.$$

将其写成 (13.204) 的形式，有

$$B = \begin{pmatrix} 2/3 & -1/12 \\ 4/3 & 1/3 \end{pmatrix}, \quad e = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad d = \begin{pmatrix} 5/12 \\ 1/3 \end{pmatrix}.$$

(13.204)是 Y_m 的非线性方程组．对 h 加以适当的限制，应用压缩映象原理，可以证明它有唯一解．但是，若方法是 A 稳定的，希望用大步长 h 进行计算．

由等式

$$\bar{Y}_m = hB\bar{F}(\bar{Y}_m) + ey(t_n) + hdf(t_n, y(t_n)) + \bar{\tau}_m \quad (13.205)$$

定义局部截断误差向量

$$\bar{\tau}_m = (\tau_{n+1}, \cdots, \tau_{n+r})^T,$$

其中

$$\bar{Y}_m = (y(t_{n+1}), \cdots, y(t_{n+r}))^T,$$

$$\bar{F}(\bar{Y}_m) = (f(t_{n+1}, y(t_{n+1})), \cdots, f(t_{n+r}, y(t_{n+r})))^T.$$

我们可以证明下面的定理．

定理 13.1 设在考虑的区域中函数 f 是连续的，并且对 y 满足常数为 L 的 Lipschitz 条件．假定初值问题 (13.201) 的解 $y(t)$ 具有所需要的任意阶连续导数．另外，设对于由 (13.204) 定义的 r 块隐式单步方法存在整数 p 和 $0 < q \leqslant p$，使局部截断误差满足 $\|\bar{\tau}_m\| = O(h^{q+1})$ 和 $|\tau_{n+r}| = O(h^{p+1})$，于是方法 (13.204) 是收敛的．整体误差的阶为 h^ν，$\nu = \min\{p, q+1\}$，即对每一个 $m = 0, 1, \cdots$ 成立 $\|Y_m - \bar{Y}_m\| = O(h^\nu)$．这时方法称为 ν 阶的．这里用的向量模是最大模．

证明 按定理的假定，存在数 r_1，r_2 使对所有 $n = mr$ 和适当小的 h 有

$$\|\bar{\tau}_m\| \leqslant r_1 h^{q+1}, \quad |\tau_{n+r}| \leqslant r_2 h^{p+1}, \quad (13.206)$$

方法 (13.204) 可以看成是以步长 rh 计算 y_n，y_{n+r}，$\cdots$ 的隐式单步方法．由 (13.204)

$$y_{n+r} = y_n + rh\left[\frac{1}{r} d_r f(t_n, y_n) + \frac{1}{r}\sum_{j=1}^{r} b_{rj} f(t, y_{n+j})\right]$$

$$= y_n + rh\Phi(t_n, y_n; rh). \tag{13.207}$$

这种过程与 Butcher 处理隐式 Runge-Kutta 是类似的，其差别是在隐式 Runge-Kuttn 中，中间结果 y_{n+1}，y_{n+2}，$\cdots$，y_{n+r-1} 是不用的．而现在用它们为对应节点上解的近似值．方法的收敛性直接由 Henrici [62] 的定理 2.2 得出，唯一需要验证的是增量函数 $\Phi(\cdot, \cdot; \cdot)$ 对第二个变量的 Lipschitz 连续性．为此，首先验证 Y_m 对 y_n 的 Lipschitz 连续性．如果对 y_n^* 有

$$Y_m^* = hBF(Y_m^*) + ey_n^* + hdf(t_n, y_n^*),$$

则应用(13.204)有

$$\|Y_m - Y_m^*\| \leqslant hL\|B\|\|Y_m - Y_m^*\| + (1 + hL\|d\|)|y_n - y_n^*|$$

因此，对于适当小的 h 和常数 $\mathscr{L}$ 有

$$\|Y_m - Y_m^*\| \leqslant \mathscr{L}|y_n - y_n^*|.$$

同样的讨论可以证明

$$\|Y_m - \bar{Y}_m\| \leqslant \mathscr{L}|y_n - y(t_n)| + r_3h^{q+1}. \tag{13.208}$$

现在有

$$\begin{aligned}
&|\Phi(t, y_n; rh) - \Phi(t, y_n^*; rh)| \\
&\leqslant \frac{L}{r}[\|d\||y_n - y_n^*| + \|B\|\|Y_m - Y_m^*\|\|] \\
&\leqslant |y_n - y_n^*|\left[\frac{L}{r}(\|d\| + \mathscr{L}\|B\|)\right].
\end{aligned}$$

于是由 Henrici 定理存在 r_4，使对任何 b 和所有 t_n $(=t_0+mrh) \leqslant b$ 有 $|y_n - y(t_n)| \leqslant r_4h^p$．再由(13.208)，对于所有节点得到所需要的误差界

$$\|Y_m - \bar{Y}_m\| \leqslant h^p(\mathscr{L}r_4 + r_3).$$

对于例 13.6 中的方法，其局部截断误差向量 $\bar{\tau}_m$ 为

$$\bar{\tau}_m = \left(\frac{h^4}{24}y^{(\mathrm{IV})}(\zeta_{n+1}), \ -\frac{h^5}{90}y^{(\mathrm{V})}(\zeta_{n+2})\right)^T.$$

因此，若 $y(t)$ 是五次连续可微的，则方法将以 $O(h^4)$ 的阶收敛．

下面来研究方法(13.204)的 A 稳定性． 将(13.204)应用到数试验方程，并令 $q = \lambda h$，则我们得到

$$(I - qB)Y_m = (e + qd)y_n, \qquad (13.209)$$

为了算法(13.204)可以以任意的 $h > 0$ 应用到所有 $\mathrm{Re}\,\lambda < 0$ 的数试验方程，假定对所有 $\mathrm{Re}\,q < 0$ 的 q，矩阵 $(I - qB)$ 是可逆的. 定义 q 的多项式

$$P(q) = \det(I - qB) = \sum_{i=0}^{r} a_i q^i, \qquad (13.210)$$

$$P_k(q) = \det(B_k(q)) = \sum_{i=0}^{r} a_{ki} q^i, \quad k = 1, 2, \cdots, r, \qquad (13.211)$$

其中 $B_k(q)$ 是矩阵 $I - qB$ 的第 k 列换成向量 $e + qd$ 后得到的矩阵. 由行列式的定义，$P(q)$，$P_k(q)$ 均是次数不超过 r 的多项式，将它们正则化使有 $P(0) = P_k(0) = 1$. 由上面对矩阵 $I - qB$ 的非奇性假定推出 $P(q)$ 的所有根或 B 的所有特征值必须有非负的实部.

利用上面的定义，并对(13.209)应用 Cramer 法则，得到

$$y_{n+k} = \frac{P_k(q)}{P(q)} y_n = \left[\frac{P_k(q)}{P(q)}\right]\left[\frac{P_r(q)}{P(q)}\right]^m y_0, \quad k = 1, 2, \cdots, r, \qquad (13.212)$$

因此，当 $m \to \infty$ 时，$Y_m \to 0$ 的充分必要条件为

$$|P_r(q)/P(q)| < 1.$$

这就得到下面的定理.

定理 13.2 r 块单步方法(13.204)为 A 稳定的充分必要条件为对所有 $\mathrm{Re}\,q < 0$ 的 q 有 $|P_r(q)/P(q)| < 1$.

仍考虑例 13.6 中的方法，矩阵 B 的特征值为

$$\mu = (3 \pm i\sqrt{3})/6.$$

将这个方法应用到数试验方程，得到

$$y_{n+1} = \frac{6 - q^2}{6 - 6q + 2q^2} y_n \equiv S(q) y_n,$$

$$y_{n+2}=\frac{3+3q+q^2}{3-3q+q^2}y_n\equiv R(q)y_n$$

于是

$$y_{2m}=R(q)^m y_0,\quad y_{2m+1}=S(q)R(q)^m y_0.$$

应用 Birkhoff 和 Varga 的引理:

引理 13.3 假定 $P(z)$ 和 $Q(z)$ 均是 z 的实多项式, $Q(z)=P(-z)$, 并且 $P(z)$ 的零点具有正的实部, 则对所有 $\mathrm{Re}z<0$, 有

$$|Q(z)/P(z)|<1,$$

立即可以推得 Clippenger-Dimsdale 方法是 A 稳定的. 这个方法的阶为 4, 而按 Dahlqnist 的结果, 梯形公式是线性多步公式中最精确的 A 稳定方法.

通常可以有两种处理方法来应用定理 13.2. 一种是对已有的方法所对应的有理函数,研究它的 A 稳定性;另外一种是从对所有 $\mathrm{Re}\,q<0$ 的 q 有 $|P_r^*(q)/P^*(q)|<1$ 的多项式 $P_r^*(q)$, $P^*(q)$ 出发,构造具有 $P(q)\equiv P^*(q)$, $P_r(q)\equiv P_r^*(q)$ 的块隐式单步方法,即 A 稳定方法.

下面我们讨论根据 Newton-Cotes 型求积公式导出的一类块隐式单步方法. 由于初值问题

$$y'=f(t,y),\quad y(t_n)=y_n \tag{13.213}$$

等价于积分方程

$$y(t)=y_n+\int_{t_n}^{t} f(\tau,y(\tau))d\tau. \tag{13.214}$$

我们用与 $f(t,y(t))$ 在节点 $t_n, t_{n+1},\cdots,t_{n+r}$ 上重合的 r 次 Lagrange 插值多项式的积分来近似(13.214)中的积分项. 从而得到

$$y(t_{n+i})=y_n+\int_{t_n}^{t_{n+i}} f(t,y(t))dt$$

的近似值.这样,我们得到形式为(13.204)的公式,其中 $d_i=L_{i0}$, $b_{ik}=L_{ik}$,

$$L_{jk} = \int_0^j l_k(s)ds,\ j = 1,\ 2, \cdots, r,\ k = 0, 1, \cdots, r,$$

$$l_k(s) = \frac{s(s-1)\cdots(s-k+1)(s-k-1)\cdots(s-r)}{k(k-1)\cdots(1)(-1)\cdots(k-r)}.$$

这些公式的局部离散误差有公式

$$\tau_j = \frac{h^{r+2}}{(r+1)!} y^{(r+2)}(\xi_j) \int_0^j s(s-1)\cdots(s-r)ds$$

$$j = 1, \cdots, r, \tag{13.215}$$

对于偶数 r，可得到更强的结果

$$\tau_r = \frac{h^{r+3}}{(r+2)!} y^{(r+3)}(\xi_r) \int_0^r s^2(s-1)\cdots(s-r)ds, \tag{13.216}$$

其中 $t_n < \xi_j < t_{n+r}$. 这些公式可以按 $r+1$ 个点的 Newton-Cotes 求积公式的误差估计方法来推导.

Newton-Cotes 块方法的系数还可以有另外一种推导方式，即寻找公式 (13,204) 中的系数使公式对所有 $1 \leqslant p \leqslant r+1$ 的函数 $y(t) = t^p$ 均是精确的 (我们已要求对 $p = 0$ 是成立的). 于是 b_{jk} 和 d_j 必须满足方程组

$$d_j + \sum_{k=1}^{r} b_{jk} = j,\ j = 1, \cdots, r, \tag{13.217}$$

$$\sum_{k=0}^{r} b_{jk} k^{p-1} = \frac{1}{p} j^p,\ j = 1, \cdots, r,$$

$$p = 2, \cdots, r+1. \tag{13.218}$$

令 $\bar{S}_j = (d_j, b_{j1}, \cdots, b_{jr})^T$, $\bar{C}_j = (j, j^2/2, \cdots, j^{r+1}/(r+1))^T$ 和 V 是 Vandermonde 矩阵，则 (13.217)、(13.218) 可写成矩阵的形式

$$V^T \bar{S}_j = \bar{C}_j,\ j = 1, \cdots, r. \tag{13.219}$$

我们看到，如果要求公式 (13.204) 具有最高的代数精确度，则系数由 (13.219) 唯一确定.

由 (13.215)，Newton-Cotes r 块方法的局部精度是将其应用

到数试验方程，得到解将满足

$$y_{n+k}=\frac{P_k(q)}{P(q)}y_n=y_ne^{kq}+O(|q|^{r+2}),\ k=1,\cdots,r,\tag{13.220}$$

其中 $P(q)$、$P_k(q)$ 由 (13.210) 和 (13.211) 确定. 由这些关系式容易得到下面的引理.

引理 13.4 对于所有的 r，$k=1,\cdots,r$ 和所有使 $P(q)\neq 0$ 的 q，有

$$P_k(q)=P(q)e^{kq}+O(|q|^{r+2})\tag{13.221}$$

引理 13.5 多项式 $P_k(q)$ 的系数 a_{ki} 由公式

$$a_{ki}=\sum_{j=0}^{i}\frac{k^{i-j}}{(i-j)!}a_j,\quad \begin{matrix} i=0,1,\cdots,r\\ k=1,2,\cdots,r\end{matrix}\tag{13.222}$$

给出.

引理 13.6 多项式 $P(q)$ 的系数满足方程

$$\sum_{i=1}^{r}a_{r-i+1}\frac{k^{i-1}}{i!}=-\frac{k^r}{(r+1)!},\ k=1,2,\cdots,r.\tag{13.223}$$

引理 13.7 多项式 $P(q)$ 的系数由

$$a_i=\frac{(r-i+1)}{(r+1)!}\mu_{r-i},\ i=0,1,\cdots,r\tag{13.224}$$

给出，其中 μ_j 由

$$(t-1)(t-2)\cdots(t-r)=\sum_{j=0}^{r}\mu_jt^j\tag{13.225}$$

确定.

事实上，令

$$\beta_j=-\frac{(r+1)!}{j!}a_{r-j+1},$$

于是由 (13.223) 看到，这样定义的 $\beta_1,\beta_2,\cdots,\beta_r$ 是对 $k=1,2,\cdots,r$ 满足 $\rho(k)=k^r$ 的唯一的内插多项式

$$\rho(x)=\sum_{i=1}^{r}\beta_ix^{i-1}$$

的系数．事实上，这样的多项式为 $\rho(x)=x^r-(x-1)(x-2)\cdots(x-r)$，则得到引理．

引理 13.8 对于所有 r 和每个 $k=0,1,\cdots,\left[\frac{r}{2}\right]$，成立

$$P_k(q)=P_{r-k}(-q),$$

其中我们定义 $P_0(q)\equiv P(q)$．

由引理 13.3 和 13.8，定理 13.2 推出，如果 $P(q)$ 的零点具有正的实部，即如果 $P(-q)$ 的零点的实部均是负的，则由 Newton-Cotes 型公式确定的 r 块隐式单步公式是 A 稳定的．由于(13.224)给出 $P(-q)$ 的系数 $(-1)^i a_i$ 的显式表示，应用 Routh 算法数值地建立了下面的结果．

引理 13.9 对于 $r=1,\cdots,8$，$P(q)$ 的零点均具有正的实部，对于 $r=9,10$，存在 $P(q)$ 的零点具有负的实部．

总结上面的推导，得到下面的结果：

定理 13.3 由 Newton-Cotes 型内插公式确定的 r 块隐式单步方法为 A 稳定的充分条件是多项式 $P(-q)$ 是稳定多项式，即其零点的实部均是负的．特别对于 $r=1,2,\cdots,8$ 所对应的块方法是 A 稳定的，并且当 r 是奇数时收敛阶为 $r+1$，而对 r 为偶数时，收敛阶为 $r+2$．对于 $r=9,10$ 方法不是 A 稳定的．

下面列出 $r=1,\cdots,8$ 的公式(13.204)中的系数 d 和 B．

$$r=1 \qquad d_1=\frac{1}{2} \qquad b_{11}=\frac{1}{2},$$

$$r=2 \qquad d=\begin{pmatrix}\frac{5}{12}\\ \frac{1}{3}\end{pmatrix} \qquad B=\begin{pmatrix}\frac{8}{12} & -\frac{1}{12}\\ \frac{4}{3} & \frac{1}{3}\end{pmatrix},$$

$$r=3 \qquad d=\begin{pmatrix}\frac{9}{24}\\ \frac{1}{3}\\ \frac{3}{8}\end{pmatrix} \qquad B=\begin{pmatrix}\frac{19}{24} & -\frac{5}{24} & \frac{1}{24}\\ \frac{4}{3} & \frac{1}{3} & 0\\ \frac{9}{8} & \frac{9}{8} & \frac{3}{8}\end{pmatrix},$$

$$r = 4 \qquad d = \begin{pmatrix} \frac{251}{720} \\ \frac{29}{90} \\ \frac{27}{80} \\ \frac{14}{45} \end{pmatrix} \qquad B = \begin{pmatrix} \frac{646}{720} & -\frac{264}{720} & \frac{106}{720} & -\frac{19}{720} \\ \frac{124}{90} & \frac{24}{90} & \frac{4}{90} & -\frac{1}{90} \\ \frac{102}{80} & \frac{72}{80} & \frac{42}{80} & -\frac{3}{80} \\ \frac{64}{45} & \frac{24}{45} & \frac{64}{45} & \frac{14}{45} \end{pmatrix},$$

$$r = 5 \quad d = \begin{pmatrix} \frac{475}{1440} \\ \frac{28}{90} \\ \frac{51}{160} \\ \frac{14}{45} \\ \frac{95}{288} \end{pmatrix}$$

$$B = \begin{pmatrix} \frac{1427}{1440} & -\frac{798}{1440} & \frac{482}{1440} & -\frac{173}{1440} & \frac{27}{1440} \\ \frac{129}{90} & \frac{14}{90} & \frac{14}{90} & -\frac{6}{90} & \frac{1}{90} \\ \frac{219}{160} & \frac{114}{160} & \frac{114}{160} & -\frac{21}{160} & \frac{3}{160} \\ \frac{64}{45} & \frac{24}{45} & \frac{64}{45} & \frac{14}{45} & 0 \\ \frac{375}{288} & \frac{250}{288} & \frac{250}{288} & \frac{375}{288} & \frac{95}{288} \end{pmatrix},$$

$$r=6\qquad d=\begin{pmatrix}\frac{19087}{60480}\\[2mm] \frac{1139}{3780}\\[2mm] \frac{685}{2240}\\[2mm] \frac{286}{945}\\[2mm] \frac{3715}{12096}\\[2mm] \frac{41}{140}\end{pmatrix}\qquad B=\begin{pmatrix}\frac{65112}{60480} & -\frac{46461}{60480} & \frac{37504}{60480} & -\frac{20211}{60480} & \frac{6312}{60480} & -\frac{863}{60480}\\[2mm] \frac{5640}{3780} & \frac{33}{3780} & \frac{1328}{3780} & -\frac{807}{3780} & \frac{264}{3780} & -\frac{37}{3780}\\[2mm] \frac{3240}{2240} & \frac{1161}{2240} & \frac{2176}{2240} & -\frac{729}{2240} & \frac{216}{2240} & -\frac{29}{2240}\\[2mm] \frac{1392}{945} & \frac{384}{945} & \frac{1504}{945} & \frac{174}{945} & \frac{48}{945} & -\frac{8}{945}\\[2mm] \frac{17400}{12096} & \frac{6375}{12096} & \frac{16000}{12096} & \frac{11625}{12096} & \frac{5640}{12096} & -\frac{275}{12096}\\[2mm] \frac{216}{140} & \frac{27}{140} & \frac{272}{140} & \frac{27}{140} & \frac{216}{140} & \frac{41}{140}\end{pmatrix},$$

$$r=7 \quad d=\begin{pmatrix} \frac{36799}{120960} \\ \frac{1107}{3780} \\ \frac{1325}{4480} \\ \frac{278}{945} \\ \frac{7155}{24192} \\ \frac{41}{140} \\ \frac{5257}{17280} \end{pmatrix} \quad B=\begin{pmatrix} \frac{139849}{120960} & -\frac{121797}{120960} & \frac{123133}{120960} & -\frac{88547}{120960} & \frac{41499}{120960} & -\frac{11351}{120960} & \frac{1375}{120960} \\ \frac{5864}{3780} & -\frac{639}{3780} & \frac{2448}{3780} & -\frac{1927}{3780} & \frac{936}{3780} & -\frac{261}{3780} & \frac{32}{3780} \\ \frac{6795}{4480} & \frac{1377}{4480} & \frac{5927}{4480} & -\frac{3033}{4480} & \frac{1377}{4480} & -\frac{373}{4480} & \frac{45}{4480} \\ \frac{1448}{945} & \frac{216}{945} & \frac{1784}{945} & -\frac{106}{945} & \frac{216}{945} & -\frac{64}{945} & \frac{8}{945} \\ \frac{36725}{24192} & \frac{6975}{24192} & \frac{41625}{24192} & \frac{13625}{24192} & \frac{17055}{24192} & -\frac{2475}{24192} & \frac{275}{24192} \\ \frac{216}{140} & \frac{27}{140} & \frac{272}{140} & \frac{27}{140} & \frac{216}{140} & \frac{41}{140} & 0 \\ \frac{25039}{17280} & \frac{9261}{17280} & \frac{20923}{17280} & \frac{20923}{17280} & \frac{9261}{17280} & \frac{25039}{17280} & \frac{5257}{17280} \end{pmatrix},$$

$$
r = 8 \quad d = \begin{pmatrix} \frac{1070017}{3628800} \\ \frac{32377}{113400} \\ \frac{12881}{44800} \\ \frac{4063}{14175} \\ \frac{41705}{145152} \\ \frac{401}{1400} \\ \frac{149527}{518400} \\ \frac{3956}{14175} \end{pmatrix}, \quad B = \begin{pmatrix}
\frac{4467094}{3628800} & \frac{4604594}{3628800} & \frac{5595358}{3628800} & -\frac{5033120}{3628800} & \frac{3146338}{3628800} & -\frac{1291214}{3628800} & \frac{312874}{3628800} & -\frac{33953}{3628800} \\
\frac{182584}{113400} & -\frac{42494}{113400} & \frac{120088}{113400} & -\frac{116120}{113400} & \frac{74728}{113400} & -\frac{31154}{113400} & \frac{7624}{113400} & -\frac{833}{113400} \\
\frac{70902}{44800} & \frac{3438}{44800} & \frac{79934}{44800} & -\frac{56160}{44800} & \frac{34434}{44800} & -\frac{14062}{44800} & \frac{3402}{44800} & -\frac{369}{44800} \\
\frac{22576}{14175} & \frac{244}{14175} & \frac{32752}{14175} & -\frac{9080}{14175} & \frac{9232}{14175} & -\frac{3956}{14175} & \frac{976}{14175} & -\frac{107}{14175} \\
\frac{230150}{145152} & \frac{7550}{145152} & \frac{318350}{145152} & -\frac{4000}{145152} & \frac{170930}{145152} & -\frac{49150}{145152} & \frac{11450}{145152} & -\frac{1225}{145152} \\
\frac{2232}{1400} & \frac{18}{1400} & \frac{3224}{1400} & -\frac{360}{1400} & \frac{2664}{1400} & \frac{158}{1400} & \frac{72}{1400} & -\frac{9}{1400} \\
\frac{816634}{518400} & \frac{48706}{518400} & \frac{1085938}{518400} & \frac{54880}{518400} & \frac{736078}{518400} & \frac{522046}{518400} & \frac{223174}{518400} & -\frac{8183}{518400} \\
\frac{23552}{14175} & -\frac{3712}{14175} & \frac{41984}{14175} & -\frac{18160}{14175} & \frac{41984}{14175} & -\frac{3712}{14175} & \frac{23552}{14175} & \frac{3956}{14175}
\end{pmatrix}
$$

本 章 附 注

§1 的材料取自韩天敏的[5].

§2 是根据 Odén 的[93]编写的.

§3 主要取自 Lambert 的[68]和 Lambert, Shaw[69].

§4 的材料取自 Ракитский, устинов 和 Черноруцкий 的书 [115] 的第二章.

§5 是据 Liniger, Odeh [79] 和费景高[2]编写的.

§6 的材料取自 Shampine, Watts [104], Watts, Shampine [111] 和 Rosser [98].

参考文献

[1] 费景高，刘德贵，Stiff 常微分方程初值问题的数值解法，计算机应用与应用数学（1978），1-2，第1—40页.

[2] 费景高，一类线性隐式多步并行算法，计算机工程与科学（1980），4，第43—58页.

[3] 费景高，应用插值的数值求解常微分方程组初值问题的分解算法的收敛性和收敛阶，数值计算与计算机应用，**5**(1984)，4，第209—218页.

[4] 韩天敏，常微分方程数值解的一个方法，应用数学与计算数学，**3**(1966)，3，第187—191页.

[5] 韩天敏，刚性常微分方程初值问题的一种数值解法，中国科学（1976），1，第21—34页.

[6] 韩天敏，崔可发，刚性方程数值方法的危险性问题，计算数学，**1**(1979)，4，第331—335页.

[7] 何爱芳，关于梯形法的局部外插，计算数学，**5**(1983)，3，第326—331页.

[8] 刘德贵，控制系统计算中的一类常微分方程的变形，数值计算与计算机应用，**2**(1981)，3，第182—190页.

[9] 刘德贵，数值求解 Stiff 常微分方程组的初值问题的组合算法，计算机工程与设计（1982），2，第41—52页.

[10] 秦元勋，运动稳定性的一般问题讲义，科学出版社，1958.

[11] 孙耿，关于梯形公式的一点注记，计算数学，**1**(1979)，4，第347—353页.

[12] 汤怀民，解 Stiff 方程的某些单步方法及其稳定性质，1979年全国计算数学学会首届年会论文(未发表).

[13] 许淞庆，常微分方程稳定性理论，上海科学技术出版社，1962年.

[14] 袁兆鼎，毕德学，根轨迹计算法，应用数学与计算数学，**1**(1964)，1，第13—17页.

[15] 袁兆鼎，常微分方程数值积分中避免导数计算的一个注记，应用数学与计算数学，**2**(1965)，3，第196—204页.

[16] 袁兆鼎，有关下山法的几个问题，数值计算与计算机应用，**1**(1980)，1，第1—7页.

[17] 袁兆鼎，黄振雄，自动控制迴路数值积分中一个问题的处理，计算机工程与设计（1981），1，第39—44页.

[18] 祝楚恒，费景高，联合应用 Runge-Kutta 公式与 Adams 公式的积分方法，电子计算机动态（1963），2，第31—35页.

[19] 祝楚恒，袁兆鼎，常微分方程数值积分的计算稳定性，计算数学，**2**(1980)，1，第77—89页.

[20] C. W. 吉尔著，常微分方程初值问题的数值解法，费景高，刘德贵，高永春译，科学出版社，1978年.

[21] Alexander, R., Diagonally implicit Runge-Kutta methods for stiff ordinary differential equations, *SIAM. J. Numer. Tnal.*, **14**(1977), 6, 1006—1021.

[22] Axelsson, O., Global integration of differential equations through lobatto quadrature, *BIT*, 4(1964), 1, 69—86.

[23] Axelsson, O., A class of A-stable methods, *BIT*, **9**(1969), 2, 185—199.

[24] Axelsson, O., A note on a class of strongly A-stable methods, *BIT*, **12**(1972),

1, 1—4.

[25] Baker, G. A., Essentials of Padé, approximates, Academic Press, 1975.

[26] Bickart, T. A., An efficient solution process for implicit Runge-Kutta methods, *SIAM J. Numer. Anal.*, **14**(1977), 6, 1022—1027.

[27] Birkholf, G. and Varga, R. S., Discretization errors for well-set Cauchy problems, *Journal of Math. and Physics*, **44**(1965), 1, 1—23.

[28] Bjurel, G., Dahlquist, G.,Lindberg, B., Linde, S., Odén, L., Survey of stiff ordinary differential equations, Report NA-70. 11(1970) Dept. of Computer Science, Royal Inst. of Technol., Stockholm, Sweden.

[29] Bjurel, G., Supplement to report on modified linear multistep methods for a class of stiff ordinary differential equations, Report NA-71. 43(1971), Dept. of Computer Science, Royal Inst. of Tecchnol., Stockholm, Sweden.

[30] Bjurel, G., Modified multistep methods for the numerical solution of a class of stiff ordinary differential equations, Report NA-72. 64(1972), Dept. of Computer Science, Royal Inst. of Technol. Stockholm, Sweden.

[31] Bjurel, G., Modified linear multistep methods for a class of stiff ordinary differential equations, *BIT*, **12**(1972), 2, 142—160.

[32] Boggs, P. T., An algorithm, based on singular perturbation theory for ill-conditioned minimization problems, *SIAM J. Numer., Anal.*, **14**(1977), 6, 830—843.

[33] Butcher, J. C., Implicit Runge-Kutta processes, *Math. Comput.*, **18**(1964), 85, 50—64.

[34] Butcher, J. C., A stability property of implicity Runge-Kutta methods, *BIT*, **15**(1975), 3, 358—361.

[35] Butcher, J. C., On the implementation of implicit Runge-Kutta methods, *BIT*, **16**(1976), 2, 237—240.

[36] Butcher, J. C., On A-stable implicit Runge-Kutta methods, *BIT*, **13**(1977), 4, 375—378.

[37] Burrage, K., A special family of Runge-Kutta methods for solving stiff differential equations, *BIT*, **18**(1978), 1, 22—41.

[38] Burrage, K. and Butcher, J. C., Stability criteria for implicit Runge-Kutta methods, *SIAM J. Numer. Anal.*, **16**(1979), 1, 46—57.

[39] Burrage, K. and Butcher, J. C., Stability criteria for implicit Runge-Kutta methods, *SIAM, J. Numer. Anal.*, **16**(1979), 1, 46—57.

[40] Calahan, D. A., A stable, accurate method of numerical integration for nonlinear systems, *Proc. IEEE*, **56**(1968), p. 744.

[41] Cash, J. R., On the integration of stiff systems of ordinary differential equations using extended backward differentiation formulae, *Numer. Math.*, **34**(1980), 3, 235—246.

[42] Chipman, F. H., A-stable Runge-Kutta processes, *BIT*, **11**(1971), 4, 384—388.

[43] Chipman, F. H., The implementation of Runge-Kutta implicit processes, *BIT*, **13**(1973), 4, 391—393.

[44] Creedon, D. M., Miller, J. J. H., The stability properties of q-step backward difference schemes, *BIT*, **15**(1975), 3, 244—249.

[45] Cryer, C. W., On the instability of high order backward-difference multistep methods, *BIT*, 12(1972), 1, 17—25.

[46] Cryer, C. W., A new class of highly-stable methods: A_0-stable methods, *BIT*, 13 (1973), 2, 153—159.

[47] Dahlqnist, G., Convergence and stability in the numerical integration of ordinary differential equations, *Math. Scand.*, 4(1956), 33—53.

[48] Dahlquist, G., A special stability problem for linear multistep methods, *BIT*, 3(1963), 1, 27—43.

[49] Dahlquist, G., Error analysis for a class of methods for stiff nonlinear initial value problems, Numerical Analysis Dundee 1975, Springer Lecture Notes in Mathematics, Vol. 506, 60—74.

[50] Ehle, B. L., High order A-stable methods for the numerical solution of systems of differential equation, *BIT*, 8(1968), 3, 276—278.

[51] Ehle, B. L., On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems, University of Waterloo Dept. Applied Analysis and Computer Science, Research Rep. No. CSRR 2010(1969).

[52] Ehle, B. L., A-stable methods and Padé *Approximations to the exponential.* *SIAM J. Math. Anal.*, 4(1973), 5, 671—680.

[53] Ehle, B. L., Lawson, J. D., Generalized Runge-Kutta processes for stiff initial-value problems, *J. Inst. Maths. Applics*, 16(1975), 1, 11—21.

[54] Enright, W. H., Second derivative multistep methods for stiff ordinary differential equations, *SIAM J. Numer. Anal.*, 11(1974), 2, 321—331.

[55] Enright, W. H., Hull, T. E., Lindberg, B., Comparing numerical methods for stiff systems, of ordinary differential equations, *BIT*, 15(1975), 1, 10—48, 中译文见"计算机应用与应用数学"(1977),8,第1—14页,9,第1—18页.

[56] Gautschi, W., On inverses of vandermonde matrices and confluent vandermonde matrices, *Numer. Math.*, 4(1962), 2, 117—123.

[57] Gear, C. W., Automatic multirate methods fot ordinary differential equations, Report 1000, Dept. of Computer Science, University of Illinois at Urbana-Champaign, January 1980.

[58] Gear, C. W., The automatic integration of stiff ordinary differential equations, Information Processing 68, ed., A, J. H. Morrell, North Holland Publishing Co., 1969, 187—193.

[59] Genin, Y., A new approach to the synthesis of stiffly stable linear multistep formulae, *IEEE transactions on circuit theory*, CT-20(1973), 4, 352—360.

[60] Hall, G. and Watt, J. M., Modern numerical methods for ordinary differential equations, Clarendon Press, 1976.

[61] Haines, C. F., Implicit integration processes with error estimation for the numerical solution of differential equations, *Comput. J.*, 12(1969), 2, 183—188.

[62] Henrici, P., Discrete variable methods in ordinary differential equations, John Wiley and Sons, 1962.

[63] Henrici, P., Error propagation for difference methods, John Wiley and Sons,

1963.

[64] Jeltsch, R., Stiff stability and its relation to A_0-and A(0)-Stability, *SIAM J. Numer. Anal.*, 13(1976), 1, 8—17.

[65] Kutta, W., Beitrag zur näherungsweisen Integration totaler differential gleichungen, *Zeit. Math. Physik*, 46(1901), 435—453.

[66] Lambert, J. D., Linear multistep methods with mildly varying coefficients, *Mathematics of Computation*, 24(1970), 109, 81—93.

[67] Lambert, J. D., Computational methods in ordinary differential equations, John Wiley, 1973.

[68] Lambert, J. D., Nonlinear methods for stiff systems of ordinary differential equations, in conference on the numerical solution of differential equations, Springer-Verlag, 1974, 75—88.

[69] Lambert, J. D. and Shaw, B., On the numerical solution of y'=f(x, v) by a class of formulae based on rational approximation, *Maths. Comput.*, 19(1965), 91, 456—462.

[70] Lambert, J. D., and Sigurdsson, S. T., Multistep methods with variable matrix coefficients. *SIAM J. Numer. Anal.*, 9(1972), 4, 715—733.

[71] Lapidus, L. and Seinfeld, J. H., Numerical solution of ordinary differential equations, Academic Press, 1971.

[72] Lawson, J. D., Generalized Runge-Kutta processes for stable systems with large Lipschitz constants, *SIAM J. Numer. Anal.*, 4(1967), 3, 372—380.

[73] Lindberg, B., On the smoothing and extrapolation for the trapezoidal rule, *BIT*, 11(1971), 1, 29—52.

[74] Lindberg, B., A simple interpolation algorithm for improvement of the numerical solution of a differential equation, *SIAM J. Numer. Anal.*, 9(1972), 4, 662—668.

[75] Lindberg, B., A stiff system package based on the implicit midpoint method with smoothing and extrapolation, in stiff differential systems edited by R. A. Willoughby, Plenum Press, 1974, 201—215.

[76] Lindberg, B., On a dangerous property of methods for stiff differential equations, *BIT*, 14(1974), 4, 430—436.

[77] Liniger, W., A criteria for A-stability of linear multistep integration formulae, *Computing*, 3(1968), 3, 280—285.

[78] Liniger, W., Global accuracy and A-stability of one-and two-step integration formulae for stiff ordinary differential equations, Conference on numerical Solution of differential equations, Dundee, 1969, ed. John, Ll. Morris, Springer-Verlag, 1969, 188—193.

[79] Liniger, W. and Odén, F., A-stable, accurate averaging of multistep methods for stiff differential equation, *IBM J. Res. Develop.*, 16(1972), 4, 335—345.

[80] Liniger, W. and Willoughby, R. A., Efficient integration methods for stiff systems of ordinary differential equations, *SIAM J. Numer. Anal.*, 7(1970), 1, 47—66.

[81] Mäkelä, M. and Nevanlinna, O., Sipilä, A. H., Exponentially fitted multistep methods by generalized Hermite-Birkhoff interpolation, *BIT*, 14(1974), 4,

437—451.

[82] Miller, J. J. H., On the location of zeros of certain classes of polynomials with applications to numerical analysis, *J. Inst. Maths. Applics*, **8**(1971), 3, 397—406.

[83] Miller, J. J. H., On weak stability, stability and the type of a polynomial, Lecture Notes in Mathematics, Vol. 228, ed. John Ll. Morris, Springer-Verlag (1971), 316—320.

[84] Miller, J. J. H., Practical algorithms for finding the type of a polynomial, studies in numerical analysis, ed B'. K. D. scaife, Academic Press (1974), 253—264.

[85] Miller, J. J. H., On the type of a polynomial relative to a circle—an open problem, Optimization Techniques, ed. G. I. Marchuk, Lecture Notes in Computer Science No. 27 Springer-Verlag (1975), 394—399.

[86] Milne, W. E. and Reynolds, R. R., Stability of a numerical solution of differential equations, *J. Assoc. Comput. Mach.*, **6**(1959), 2, 196—203; Part II, *J. Assoc. Comput. Mach.*, **7**(1960), 1, 46—56.

[87] Milne, W. E. and Reynolods, R. R., Fifth-order methods for the numerical solution of ordinary differential equations, *J. Assoc. Comput. Mach.*, **9**(1962), 1, 64—70.

[88] Miranker, W. L., Numerical methods of boundary layer type for stiff systems of differential equations, *Computing*, **11**(1973), 2, 221—234.

[89] Nørsett, S. P., A criterion for A(α)-stability of linear multistep methods, *BIT*, **9**(1969), 3, 259—263.

[90] Nørsett, S. P., Multiple Padé, approximations to the exponential function, Mathematics Department University of Trondheim, Report, No 4/74.

[91] Nørsett, S. P., C-polynomials for rational approximation to the exponential function, *Numer. Math.* **25**(1975), 1, 39—56.

[92] Nørsett, S. P., Runge-Kutta methods with a multiple real eigenvalue only, *BIT*, **16**(1976), 4, 388—393.

[93] Odén, L., An experimental and theoretical analysis of the SAPS-method for stiff ordinary differential equations, Report NA 71, 28(1971), Dept of Inf. Proc, The Roy. Inst. Tech., stockholm, Sweden.

[94] Oppelstrup, J., One-step methods with interpolation for Volterra functional differential equation, TRITA-NA-7623(1976), Roy Inst. Tech., Stockholm, Sweden.

[95] Prothero, A. and Robinson, A., On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations, *Maths. Comput.*, **28**(1974), 125, 145—162.

[96] Ralston, A., A first course in numerical analysis, McGraw-Hill, 1965.

[97] Rosenbrock, H. H., Some general implicit processes for the numerical solution of differential equations, *Comput. J.* **5**(1963), 4, 329—330.

[98] Rosser, J., A Runge-Kutta for all seasons, *SIAM Rev.*, **9**(1967), 4, 417—452.

[99] Sacks-Davis, R., Solution of stiff ordinary differential equations by a second derivative method, *SIAM J. Numer. Anal.*, **14**(1977), 6, 1088—1100.

[100] Saff, E. B. and Varga, R. S., On the zeros and poles of pade approximats to e^x, *Numer. Math.*, 25(1975), 1, 1—4.

[101] Saff, E. B. and Varga, R. S., Zero-free parabolic region for sequences of polynomials, *SIAM Journal on Mathematical Analysis*, 7(1976), 3, 344—357.

[102] Sarkany, E. F. and Liniger, W., Exponential fititng of matricial multistep methods for ordinary differential equations, *Maths. Comput.*, 28(1974), 128, 1035—1052.

[103] Shampine, L. F. and Gordon, M. K., Computer Solution of Ordinary differential equations, San Francisco: W. H. Free man, 1975.

[104] Shampine, L. F. and Watts, H. A., Block implicit one-step methods, *Maths. Comput.*, 23(1969), 108, 731—740.

[105] Söderlind, G., A note on the stability region of the backwards differentiation methods, Report. TRITA-NA-7708 (1977), Roy. Inst. Tech., Stockholm, sweden.

[106] Stetter, H. J., Analysis of discretization methods for ordinary differential equations, Springer-Verlag, 1973.

[107] Strang, G., Accurate partial difference methods, II: Nonlinear problems, *Numer. Math.*, 6(1964), 1, 37—46.

[108] Uspensky, J. V., Theory of equations, McGraw-Hill, 1948.

[109] Varan, J. M., Stiffly stable linear multistep methods of extended order, *SIAM J. Numer. Anal.*, 15(1978), 6, 1234—1246.

[110] Wanner, G. and Hairer, E., Nørsett, P., Order stars and stability theorems, *BIT*, 18(1978), 4, 475—489.

[111] Watts, H. A. and Shampine, L. F., A- stable block implicit one-step methods, *BIT*, 12(1972), 3, 252—266.

[112] Wells, D. R., Multirate linear multistep methods for the solution of systems of ordinary differential equations, Report No, UIUCDCS-R-82-1093, Dept. Computer Sci., Illinois, 1982.

[113] Widlund, O. B., A note on unconditionally stable linear multistep methods, *BIT*, 7(1967), 1, 65—70.

[114] Васильева, А. Б., Бутузов, В. Ф., Асимптотические разложения решений сингулярно возмущенных уравнений, 'наука' 1973.

[115] Ракитский, Ю. В., Устинов, С. М., Черноруцкий, И. Г., Численные методы решения жестких систем, наука 1979.

[ＧｅｎｅｒａｌＩｎｆｏｒｍａｔｉｏｎ]
书名＝刚性常微分方程初值问题的数值解法
作者＝袁兆鼎　　费景高　　刘德贵
页数＝４９４
ＳＳ号＝１０２３７０１８
出版日期＝１９８７年１１月第１版